MODEL SELECTION FOR GRAPHICAL MARKOV MODELS

ONG MENG HWEE, VICTOR

NATIONAL UNIVERSITY OF SINGAPORE

2014

MODEL SELECTION FOR GRAPHICAL MARKOV MODELS

ONG MENG HWEE, VICTOR

(B.Sc. National University of Singapore)

A THESIS SUBMITTED

FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

DEPARTMENT OF STATISTICS AND APPLIED PROBABILITY

NATIONAL UNIVERSITY OF SINGAPORE

2014

ACKNOWLEDGEMENTS

First and foremost, I would like to express my deepest gratitude to my supervisor, Associate Professor Sanjay Chaudhuri. He has seen me through all of my four and a half years as a graduate student, from the initial conceptual stage and through ongoing advice to the end of my PhD. I am truly grateful for the tremendous amount of time he put aside and support he gave me. Furthermore, I want to thank him for encouraging me to do PhD studies as well as introducing me to the topic of graphical model selection. This dissertation would not have been possible without his help.

I am grateful to Professor Loh Wei Liem for all his invaluable advice and encouragement. I also would like to thank Associate Professor Berwin Turlach, also one of the co-authors for the paper "Edge Selection for Undirected Graph", for his guidance.

I want to thank all my friends, seniors and the staffs in Department of Statistics and Applied Probability who motivated and saw me through all these years. I also would like to thank Ms Su Kyi Win, Ms Yvonne Chow and Mr Zhang Rong for their support.

I wish to thank my parents for their undivided support and care. I am grateful that they are always there when I need them. Last but not least, I would like to thank my fiancée, Xie Xueling, for her support, love and understanding.

CONTENTS

SUMMARY

Model selection has generate an immense amount of interest in Statistics. In this thesis, we investigate methods for model selection for the class of Graphical Markov models. This thesis is split into three parts.

In the first part (Chapter 4), we look at model selection for undirected graphs. Undirected graphs provide a framework to represent relationships between variables. It has seen many applications, like genetic networks etc. We develop an efficient method to select the edges of an undirected graph. Based on group LARS, our method combines the computational efficiency of LARS and the ability to force the algorithm to always select a symmetric adjacency matrix for the graph. Properties of 'Edge selection' method are studied. We further apply our method on the isoprenoid pathways in Arabidopsis thaliana data set.

Most penalized likelihood based method penalizes all parameters in a model. In many applications encountered in real life, some information about the underlying model is known. In the second part (Chapter 5), we consider a LASSO based penalization method when the model is partially known. We consider conditions for selection consistency of such models. It is seen that these consistency conditions are different from the corresponding conditions when the model is completely unknown. In fact, our study reveals

that in many cases, knowing the model partially may not always help in selection consistency.

In the third part (Chapter 6), we develop results that can uniquely construct a graph from available information about partial regression coefficients among vertices. In particular, we look at some "almost qualitative" inequalities among signed partial correlation and regression coefficients between the vertices on a graph. General results for Gaussian tree models and polytree models are obtained. We also show how these methods can identify single factor model from a given dataset.

List of Abbreviations

AR	Autoregressive
CPG	Cocktail party graph
CV	Cross Validation
DAG	Directed acyclic graph
ES	Edge Selection
IPF	Iterative proportional fitting
LARS	Least angle regression
LASSO	Least Absolute Shrinkage Selection Operator
OLS	Ordinary Least Squares
MB	Neighborhood selection approach by Meinshausen and Buhlmann
PLARS	Partial least angle regression
PLASSO	Partial LASSO
SPACE	Partial Correlation Estimation by Joint Sparse Regression Models
UG	Undirected graph

List of Figures

List of Tables

CHAPTER 1
Introduction

1.1 Introduction

Many real world applications of statistics involve studying variables which may interact and depend on each other. The problem of model selection is one of the primary problems in statistics and has huge potential for many applications. For a practitioner, model selection procedures provide empirical evidence about the underlying models and by that help in studying natural phenomena.

Model selection poses many conceptual and implementational difficulties. The number of possible models are exponential in terms of the number of auxiliary variables. Thus, when the number of variables are large, computing the loss function for each of these models is impossible. Moreover, models with more variables usually explain more variation in the data, and can result in over fitting. So methods which penalize against larger models are used. However, these methods may require us to search all the models and in some cases the amount of penalization required has to be estimated.

In recent years, various LASSO [Tibshirani, 1996] based methods have become very popular in model selection problems. These methods select a model by using penalization to shrink regression coefficients to zero. Furthermore, these methods do not require

computation of all the models in the model space. Algorithms which allow fast computation exist [Friedman et al., 2007, Efron et al., 2004, Osborne et al., 2000]. It is also shown that under certain conditions, these methods will asymptotically choose the correct model.

Graphical Markov models [Lauritzen, 1996, Whittaker, 1990] use various graphs to represent interactions between variables in a stochastic model. Furthermore, they provide an efficient way to study and represent multivariate statistical models. Nodes in the graph are assumed to represent usually univariate random variables and the pattern of the edges represent conditional or unconditional independence relationships between them. The aim of a graphical Markov model is to provide a representation so that these interactions can be read off from the graph merely by eye estimation. In fact, the insight these patterns provide is very useful in understanding complex relationships. The examples of such graphical models abound. They have been used in gene networks, gene pathways, speech recognition, machine learning, environmental statistics, etc.

Model selection for Graphical Markov Models is interesting as the set of possible graphical Markov models can be huge, and thus it is impossible to evaluate all possible models. In this thesis, we study various approaches of model selection for graphical Markov models. We first need to specify what kind of graph we are selecting. This is usually specified by the background knowledge of the problem. Our focus is on the model selection of two types of graph, undirected graph (UG) and directed acyclic graph (DAG).

1.2 Outline of thesis

In Chapter 2 and 3, we introduce definitions and basic terminologies for Gaussian graphical models and LASSO. A basic literature review is also conducted, which provides the foundation for the rest of the chapters.

In Chapter 4, we look into a new method of model selection for undirected graphs, which is based on linear regression but does not suffer from the problem of asymmetric selection. Our method is based on group LARS [Yuan and Lin, 2006]. Due to the

linearity inherited from LARS, this algorithm provides a quick and efficient method to select an undirected graph. Properties of this 'Edge selection' method are explored both analytically as well as through simulation study. We also apply our method on the isoprenoid pathways in Arabidopsis thaliana data set.

In Chapter 5, we consider the situation where some of the coefficient are already known. In standard LASSO, it is usually assumed that a model is completely unknown. Using the weighted LASSO [Zou, 2006], we observe that we can remove the penalization on some of the coefficient estimates by setting some of the weights to be exactly zero. We found that this affects the optimization problem and its asymptotic properties. A detailed asymptotic study of the necessary and sufficient conditions required for selection consistency is conducted.

Each graph uniquely specifies and represents a set of conditional independence relationships between its vertices. The opposite assertion is not always true. It turns out that only conditional independence relations do not completely specify a graphical model. Some knowledge about non zero partial correlations is also required. Chaudhuri and Richardson [2003] study information inequalities on directed acyclic graphs. Similar comparisons of absolute partial regression coefficients are possible [Chaudhuri and Tan, 2010]. In chapter 6, we extend these results to make comparisons among signed partial correlations, which are relevant to model selection.

CHAPTER 2

LASSO

2.1 LASSO for linear Regression

Suppose we are given a response vector $\mathbf{Y}$ where

$$\mathbf{Y} = (Y_1, ..., Y_n)^T$$

and a matrix of covariates $\mathbf{X}$ where

$$\mathbf{X} = \begin{pmatrix} X_{11} & X_{12} & \dots & X_{1p} \\ X_{21} & X_{22} & \dots & X_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ X_{n1} & X_{n2} & \dots & X_{np} \end{pmatrix} = (\mathbf{x_1^T}, ..., \mathbf{x_n^T})^\mathbf{T} = (\mathbf{X_1}, ..., \mathbf{X_p})$$

and

$$\mathbf{X}_j = (X_{1j}, ..., X_{nj})^T, \mathbf{x}_i = (X_{i1}, ..., X_{ip}).$$

Without loss of generality, we assume that $\mathbf{Y}$ is centered and the columns of $\mathbf{X}$ are

standardized such that

$$\sum_{i=1}^{n} x_{ij}^2 = 1, \ \sum_{i=1}^{n} x_{ij} = 0, \ \sum_{i=1}^{n} Y_i = 0.$$

This would imply that the regression model can be expressed as

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{2.1.1}$$

where $\boldsymbol{\epsilon}$ is a vector of errors which are normally distributed with mean $\mathbf{0}$ and variance $\sigma^2\mathbf{I}_p$. Note that each entry of $\mathbf{Y}$ can be expressed as

$$Y_i = \beta_1 x_{1i} + ... + \beta_p x_{pi} + \epsilon_i = \mathbf{x_i}\boldsymbol{\beta} + \boldsymbol{\epsilon_i}$$

for $1 \leq i \leq n$.

In a real data application, it is often seen that the true model depends only on a few of the available predictors. That is, $\beta_j = 0$ for a vast number of predictors $\mathbf{X}_j$. It is well known that the coefficients estimated by minimizing residual squared errors (Ordinary least square(OLS)) estimates will not produce a parsimonious model.

There are several difficulties in using OLS estimates in presence of vast number of predictors. The fitted model may be difficult to interpret. The bias and variance of OLS estimates depend on the specific model. As for example, the OLS estimator is unbiased when it is over-specified and is biased and inconsistent when the model is underspecified. Moreover, even if the OLS estimate is unbiased, their variances may be large and this may cause the corresponding predictors to be inaccurate.

An alternative to minimizing the residual square errors is the bridge estimator [Frank and Friedman, 1993]. In particular, it estimates $\hat{\boldsymbol{\beta}}$ by solving the following equation

$$\hat{\boldsymbol{\beta}} = argmin_{\beta}||\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}||_2^2 + \lambda_n \sum_{j=1}^{p} |\beta_j|^r, \tag{2.1.2}$$

where r and λ are positive real numbers that are selected or determined before solving

equation (2.1.2). Larger values of λ shrink the total sum of the absolute value of β_j, i.e. $\sum_{j=1}^{p} |\beta_j|^r$, to a smaller value. The value of r determines the shape of the shrinkage function. When $r = 2$, the procedure is called ridge regression, which has a larger bias and a smaller variance compared to the OLS estimates. Unfortunately, similar to the OLS procedure, the ridge regression is not able to perform variable selection. This is because β_j^2 is differentiable everywhere, and therefore ridge regression does not shrink the estimates to zero fast enough [Hastie et al., 2009].

It is known that the bridge estimator would produce estimates that are exactly zero if $r \leq 1$ [Knight and Fu, 2000, Linhart and Zucchini, 1986]. Notice that when r is strictly less than one, the penalty function is not convex anymore. So the case when $r = 1$ combines two properties. The first being that it can shrink some estimates to zero. On the other hand, the penalty function is still convex. Therefore, one can use convex optimization techniques to numerically calculate the estimates. The bridge regression with $r = 1$ is called LASSO, which was first proposed by Tibshirani [1996]. Using the convexity of the LASSO problem, several existing convex optimization methods have been used to solve (2.1.2). Examples of such algorithm are Least angle regression (LARS) [Efron et al., 2004] and homotopy algorithm [Osborne et al., 2000]. These two algorithms produce the whole solution path of LASSO with varying values of λ. For a specified λ, approximation method such as pathwise coordinate descent method [Friedman et al., 2007] is also available.

Another advantage of using LASSO is that it does not require one to search for the whole model space, which can be extremely large. This is specially true for graphical Markov models where this model space is huge.

2.2 Asymptotics of LASSO

Estimation consistency of LASSO has been studied by Knight and Fu [2000] under some regularity conditions. In particular, the following regularity conditions are assumed:

$$\frac{1}{n}\mathbf{X}^T\mathbf{X} = \mathbf{C} \to \Sigma, \; as \; n \to \infty \tag{2.2.1}$$

where $\mathbf{C}$ is a positive definite matrix, and

$$\frac{1}{n} max_{1 \leq i \leq n} \mathbf{x_i} \mathbf{x_i}^T \to 0, \; as \; n \to \infty. \tag{2.2.2}$$

Regularity conditions (2.2.1) and (2.2.2) are known to be rather weak, and holds if each $\mathbf{x}_i$ are identically and independently distributed with finite second order moments [Knight and Fu, 2000].

Define the LASSO estimator as $\hat{\boldsymbol{\beta}}^{LASSO}$ where $\hat{\boldsymbol{\beta}}^{LASSO}$ is estimated as

$$\hat{\boldsymbol{\beta}}^{LASSO} = argmin_{\beta} ||\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}||_2^2 + \lambda \sum_{j=1}^{p} |\beta_j|. \tag{2.2.3}$$

Also, we define $sign(\boldsymbol{\beta})$ as a vector with entries $sign(\beta_1), ..., sign(\beta_p)$, where

$$sign(\beta_j) = \begin{cases} 1 & \beta_j > 0, \\ -1 & \beta_j < 0, \\ 0 & \beta_j = 0. \end{cases}$$

Knight and Fu [2000] show consistency of LASSO under two different rates of λ_n, namely when $\lambda_n = o(n)$ and $\lambda_n = o(\sqrt{n})$. Their results are reproduced below.

Theorem 2.1 *Under regularity conditions* (2.2.1) *and* (2.2.2) *and* $\mathbf{C}$ *is nonsingular,*

(1) If $\frac{\lambda}{n} \to \lambda_0 \geq 0$, *then*

$$\hat{\boldsymbol{\beta}}^{LASSO} \to_p argmin(V_1)$$

where

$$V_1(\mathbf{u}) = (\mathbf{u} - \boldsymbol{\beta})^T \Sigma (\mathbf{u} - \boldsymbol{\beta}) + \lambda_0 \sum_{j=1}^{p} |u_j|.$$

(2) If $\frac{\lambda}{\sqrt{n}} \to \lambda_0 \geq 0$, *then*

$$\sqrt{n}(\hat{\boldsymbol{\beta}}^{LASSO} - \boldsymbol{\beta}) \to_d argmin(V_2)$$

where

$$V_2(\mathbf{u}) = -2\mathbf{u}^T\mathcal{W} + \mathbf{u}^T\Sigma\mathbf{u} + \lambda_0 \sum_{j=1}^{a_3} \left[u_j sign(\beta_j) I(\beta_j \neq 0) + |u_j| I(\beta_j = 0)\right].$$

A few conclusions can be drawn from Theorem 2.1 above. First, $\lambda_n/n \to 0$ implies that $\hat{\boldsymbol{\beta}}^{LASSO}$ is unbiased and therefore ensures estimation consistency. Second, when λ_n is of order $\sqrt{n}$, $\hat{\boldsymbol{\beta}}^{LASSO}$ is asymptotically convergent in distribution but biased. The third conclusion is on selection consistency. We say that a selected model is consistent in selection if $\beta_j = 0$ whenever $\hat{\beta}_j = 0$ and $\beta_j \neq 0$ whenever $\hat{\beta}_j \neq 0$. In fact, Zou [2006] deduced from the second part of Theorem 2.1 that the LASSO problem is not asymptotically selection consistent with positive probability when λ_n is of the order $o(\sqrt{n})$.

Therefore, in order for LASSO to be consistent in selection, we should consider the case when $\lambda_n/\sqrt{n} \to \infty$. In fact, Zhao and Yu [2006] considered the case when $\frac{\lambda_n}{\sqrt{n}} \to \infty$ and $\frac{\lambda_n}{n} \to 0$. They prove that under these conditions, there exist Irrepresentable conditions, which are sufficient and necessary for sign consistency for fininte p. In here, sign consistency holds when $sign(\hat{\boldsymbol{\beta}}^{LASSO}) = sign(\boldsymbol{\beta})$. Note that Sign consistency is stronger than selection consistency because the latter only requires the zeroes to be matched.

2.3 Extensions of LASSO

Since the penalized least square and penalized likelihood based methods have been proven to be extremely useful in model selection and dimension reduction. Several extensions of LASSO have been proposed in the literature. We specifically consider the weighted lasso [Zou, 2006] and group LASSO [Yuan and Lin, 2006] below. These procedures are useful in graphical model selection.

2.3.1 Weighted LASSO

In many real application, it is possible to specify a relative degree of importance of the predictors in the model. In such cases, it is desirable that the different coefficients β_j are shrunk by different amount. Standard LASSO is not capable of doing that. In that situation, the weighted LASSO [Zou, 2006] can be used. Weighted LASSO estimates β as

$$\hat{\boldsymbol{\beta}} = arg\min_{\beta} ||\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}||_2^2 + \lambda \sum_{j=1}^{p} w_j|\beta_j|. \tag{2.3.1}$$

The main difference between the standard LASSO problem and weighted LASSO problem in (2.3.1) are the weights that are added to the penalty function. It is clear that assigning a smaller value of w_j would imply that the corresponding β_j would not be as heavily penalized as the others.

The estimate $\hat{\boldsymbol{\beta}}$ can be easily obtained by modifying the existing LASSO algorithm. In fact, if $w_j \neq 0$, the solution of (2.3.1) can be obtained from the reformulated LASSO problem,

$$\hat{\boldsymbol{\beta}} = arg\min_{\beta} ||\mathbf{Y} - \mathbf{X}^*\boldsymbol{\beta}||_2^2 + \lambda \sum_{j=1}^{p} |\beta_j|,$$

where $\mathbf{X}_j^* = \mathbf{X}_j/w_j$.

The adaptive LASSO, introduced by Zou [2006], is a special case of the weighted LASSO. Here, the weights are taken to be, $w_j = |\beta_j^{ols}|^{-1}$, where β_j^{ols} is the ordinary least square estimate from the full model. It is clear that a relatively large value of $|\beta_j^{ols}|$ would result in a smaller weight, which in turn would imply a weaker penalization of β_j. It was shown [Zou, 2006] under reasonable conditions on λ, the adaptive LASSO is consistent even when the standard LASSO is not.

2.3.2 Group LASSO

In standard LASSO, we select variables based on their individual strength and influence on the model. This is undesirable when the variables are interpretable only when

they are part of a group of variables. Yuan and Lin [2006] show several examples of such variables in multi-factor analysis-of-variance(ANOVA) and additive models with polynomial or nonparametric components. As for example, second order interactions are interpretable only in the presence of main effects. Thus, a variable selection procedure should include second order interactions only when the main effects are in the model.

The Group LASSO procedure selects groups of variables instead of individual ones. In this procedure, other then putting the variables in groups, the penalty function is modified to penalize the whole groups.

For that purpose, the p columns in $\mathbf{X}$ are first divided into K different subgroups. That is, the new data matrix looks like $\mathfrak{X} = [\mathfrak{X}_1, ..., \mathfrak{X}_K]$, which is a permutation of the columns of $\mathbf{X}$, i.e. $\mathbf{X} = \mathbf{P}\,[\mathfrak{X}_1, ..., \mathfrak{X}_K]$ for some permutation matrix $\mathbf{P}$. Re-expressing $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ as

$$\mathbf{Y} = \sum_{J=1}^{K} \mathfrak{X}_J \mathfrak{B}_J + \boldsymbol{\epsilon},$$

Yuan and Lin [2006] proposed a group LASSO problem which estimates $\hat{\boldsymbol{\beta}}$ as

$$\hat{\boldsymbol{\beta}} = argmin_{\beta} ||\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}||_2^2 + \lambda \sum_{J=1}^{K} ||\mathfrak{B}_J||_{\mathfrak{B}_J},$$

where

$$||\mathfrak{B}_J||_{\mathfrak{B}_J} = \sqrt{\mathfrak{B}_J^T \mathcal{K}_J \mathfrak{B}_J}$$

and $\mathcal{K}_J$ is pre-defined symmetric positive definite matrix. A common choice of $\mathcal{K}_J$ is the identity matrix. Additionally, it is often assumed that the columns of $\mathbf{X}_J$ are orthonormal for each J. This happens by construction in ANOVA. For more general structure, Gram-Schmidt orthonormalization may be used.

Using numerous simulation studies, Yuan and Lin [2006] showed that group LASSO has good performance over traditional methods such as stepwise backward elimination, especially in problems such as ANOVA. However, the solution path of group LASSO is non-linear which makes it computationally intensive.

2.4 LARS

Least angle regression (LARS), introduced by Efron et al. [2004], is a geometric way of solving the LASSO problem. It is an efficient algorithm to produce a complete solution path for LASSO penalization.

Let $\hat{\mathbf{r}} = \mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}$ be the residual vector, where $\hat{\boldsymbol{\beta}}$ is the current estimate of the coefficient, LARS selects the model by including the variables which has the highest association with the current residual vector, i.e. the association of $\mathbf{X}_j$ and $\hat{\mathbf{r}}$ is defined as $|\mathbf{X}_j^T\hat{\mathbf{r}}|$.

The algorithm proceeds as follows.

(1) **[Initialization.]** At step 0, we start with $\hat{\boldsymbol{\beta}} = 0$. Therefore, $\hat{\mathbf{r}} = \mathbf{Y}$. LARS picks a predictor, say $\mathbf{X}_{j_0}$, which has the highest association with the response vector, i.e. $|\mathbf{X}_{j_0}^T\mathbf{Y}| > |\mathbf{X}_j^T\mathbf{Y}|$ for any $j \in \{1, ..., p\}, j \neq j_1$. We denote the active set $\mathcal{E}$ as the set that contains variables that is selected by LARS. Thus, $j_0 \in \mathcal{E}$.

(2) **[Initial Direction.]** LARS then moves $\hat{\boldsymbol{\mu}} = \mathbf{X}\hat{\boldsymbol{\beta}}$ in the direction of the projection of $\mathbf{Y}$ on $\mathbf{X}_{j_0}$ until some other variable, say $\mathbf{X}_{j_1}$ has as much association as $\mathbf{X}_{j_0}$ with the residual vector $\hat{\mathbf{r}}$. At this point, the active set $\mathcal{E}$ includes j_0 and j_1. Let $k = 1$.

(3) **[Direction Change.]** At step k, LARS changes direction, and $\hat{\boldsymbol{\mu}}$ moves in a direction that is equiangular to all the predictors in the active set.

(4) **[Point of Direction Change.]** LARS moves towards the direction stated above until either one of these three things occur.

   (a) **[Selection Rule.]** Another variable, say $\mathbf{X}_{j_{k+1}}$, has as much association with the variables in the active set.

   (b) **[Dropping Rule.]** One of the coefficient estimate, say $\hat{\beta}_{j_{k+1}}$, in the active set becomes zero.

   (c) **[Stopping Rule.]** $\mathbf{X}^T\hat{\mathbf{r}}$ is equals to zero.

   Step $k = k + 1$. If (a) happens, add j_{k+1} to $\mathcal{E}$ and go back to (3). If (b) happens, drop j_{k+1} from $\mathcal{E}$ and go back to (3). If (c) happens, the algorithm ends.

It is shown Efron et al. [2004] that the solution path of the above algorithm is equivalent to the full LASSO solution.

2.4.1 Group LARS

The group LARS [Yuan and Lin, 2006] is an extension of the LARS method proposed by Efron et al. [2004]. Group LARS selects spaces spanned by $\mathfrak{X}_J$, instead of individual variables. The degree of association between the residual vector and the space spanned by $\mathfrak{X}_J$ can be defined through the angle between the residual vector and its projection on that space. Using this degree of association, an adaption of the LARS algorithm is proposed to select group $\mathfrak{X}_J$. In particular, in order to add a group, say $\mathfrak{X}_{J_2}$, when $\mathfrak{X}_{J_1}$ is already in the model, we require $||\mathfrak{X}_{J_1}^T \hat{\mathbf{r}}||^2 = ||\mathfrak{X}_{J_2}^T \hat{\mathbf{r}}||^2$. This procedure is continued until $\mathfrak{X}^T \hat{\mathbf{r}} = 0$.

If the whole matrix $\mathbf{X}$ is orthogonal, which happens for ANOVA. It can be seen [Yuan and Lin, 2006] that group LASSO and group LARS are equivalent. We use group LARS type procedure for selecting undirected graph. The group wise selection allows us to keep the adjacency matrix symmetric. The LARS procedure provides a computationally efficient way to inspect the whole path. The details are described in Chapter 4.

2.5 Multi-fold cross validation

The tuning parameter λ in the LASSO problem controls the amount of regularization. A good choice of λ would select a model that is close to the true model with good prediction accuracy. However, it is difficult to check if a particular value of λ selects a model that is close to a true model. Therefore, it is often that only prediction accuracy is considered. In linear regression, the most common measurement used is the residual sum of squares.

In multi-fold cross validation, we split our dataset into B different groups, and allocate each group into either the training data or the test data. We consider the situation where only one group is used for the test data while the rest is allocated to the training data. Therefore, there are B different ways to split these groups.

In other words, we randomly split the rows of data matrix $\mathbf{X}$ and $\mathbf{Y}$ are into B different sets, $\mathbf{X}_1^\star, \ldots, \mathbf{X}_B^\star$ and $\mathbf{Y}_1^\star, \ldots, \mathbf{Y}_B^\star$, where each $\mathbf{Y}_b^\star$ is of size n_b. For any $b = 1, 2, \ldots, B$, let $\mathbf{X}_{-b}^\star$ and $\mathbf{Y}_{-b}^\star$ be the data matrix and response vector obtained after removing $\mathbf{X}_b^\star$ and $\mathbf{Y}_b^\star$ respectively. For any nonnegative λ, let $\hat{\boldsymbol{\beta}}_{-b}^*(\lambda)$ be the coefficient estimate obtained from equation (2.2.3), based on $\mathbf{Y}_{-b}^\star$ and matrix $\mathbf{X}_{-b}^\star$. Define

$$R_b(\lambda) = || \mathbf{Y}_b^\star - \mathbf{X}_b^\star \hat{\boldsymbol{\beta}}_{-b}^*(\lambda) ||_2^2 / n_b, \tag{2.5.1}$$

$$\bar{R}(\lambda) = \sum_{b=1}^{B} R_b(m)/B. \tag{2.5.2}$$

We pick λ which minimizes $\bar{R}(\lambda)$.

Note that multi-fold cross validation can also be extended to group LARS type procedure for selecting undirected graphs. The details can be found in Chapter 4.

CHAPTER 3

Graphical models

A graph G is defined as a pair $G = (V, E)$ where $V = \{1, ..., p\}$ is the set of vertices or nodes and $E \subset V \times V$ is the set of edges. In our discussion, each vertex $i \in 1, .., p$ in the graph would represent an univariate $\mathbf{X}_i$. For i, j and k, we say that vertex i is independent of vertex j given vertex k, $i \perp\!\!\!\perp j|k$, if and only if $\mathbf{X}_i \perp\!\!\!\perp \mathbf{X}_j|\mathbf{X}_k$. Similarly, i is said to be independent of j if $\mathbf{X}_i \perp\!\!\!\perp \mathbf{X}_j$.

We consider two types of edges for our graphs. In particular, we have the following definition.

Definition 3.1 *Let $G = (V, E)$ be a graph, where $V = \{1, ..., p\}$ is the set of vertices or nodes and $E \subset V \times V$ is the set of edges. If*

(1) both (t, j) and (j, t) are in set E, then there is an undirected edge between vertex t and j.

(2) $(t, j) \in E$ and $(j, t) \notin E$, then there is a directed edge from vertex t to j.

(3) both (t, j) and (j, t) is not in set E,, then there is no edge between vertex t and j.

Note that an undirected edge is represented by a straight line while a directed edge from vertex t to j is represented by an arrow pointing to j.

Examples of undirected graph(UG) include Markov random field, concentration Graph, phylogenetic trees etc. They are also used to represent a genetic networks or a social network. Directed ayclic graph(DAG) are sometimes called Bayesian networks. They have been used in pedigree analysis, hidden Markov models, spatieo temporal models, genetic pathways and other various models of causes and effects.

In graphical model selection, our interest is in selecting the edges of a graph. We concentrate on UG and DAG. We review some notions in graphical Markov models and some available methods for undirected and directed acyclic graph selection.

3.1 Undirected Graphs

As the name suggests, undirected graphs are graphs with only undirected edges. Before describing the Markov properties, we need to define the notation of a path between two vertices on the graph.

Definition 3.2 *Let $G = (V, E)$ be an undirected graph. For two distinct vertices a and c in V. A path π of length k is a set of k non-repeating vertices $v_1, ..., v_k$ such that $a = v_1$, $c = v_k$, and for every i from $1, ..., k-1$, $(v_i, v_{i+1}) \in E$ and $(v_{i+1}, v_i) \in E$.*

Note that by our definition, the endpoints a and c are also on the path π. There may be more than one path between two vertices a and c in G. If G is a tree or a forest, then the path between two connected vertices a and c is unique.

3.1.1 Markov properties represented by an undirected graph

Several list of conditional independence relationships could be constructed from an undirected graph. Not all of such list are equivalent. One important list is called the global Markov property.

Definition 3.3 (Separation) *Let A, C and S be three disjoint sets of V (S can be empty set). Then, we say that S separates A from C if for any node $a \in A$ and $c \in C$ and any path π between a and c, there exist a vertex $s \in S$ such that $s \in \pi$.*

An undirected graph $G = (V, E)$ is said to obey the global Markov property if for disjoint subsets A, B and S in V (S may be empty), S separates A from B in G implies $A \perp\!\!\!\perp B|S$. The global Markov property is the largest listing of conditional independence relations for a graph. All other such list (eg. local, pairwise properties etc) are contained in it. For details, we follow Lauritzen [1996] and Whittaker [1990].

The pairwise Markov property is relevant for Gaussian parameterization of undirected graph which we next define. An undirected graph $G = (V, E)$ is said to obey pairwise Markov property if for all $1 \leq t, j \leq p$, if there is no undirected edge between node t and j, then $t \perp\!\!\!\perp j|\mathbf{p}\backslash\{t, j\}$.

For any undirected graph, the global Markov property implies the pairwise Markov property. The opposite implication is in general false. However, if the joint distribution of the vertices is Gaussian, then the pairwise and global Markov property are equivalent. Furthermore, for Gaussian distribution, if there is no edge between j and t, the corresponding entry in the inverse covariance matrix is zero,

This fact is exploited in the parameterization of Gaussian undirected graph and forms the backbone of any model selection procedure for these graphs.

3.1.2 Parameterization

Suppose $\mathbf{X}$ is a $n \times p$ data matrix, where each row follows a multivariate normal distribution with positive definite covariance matrix Σ. We denote the (i, j) entry of Σ as $\Sigma_{i,j}$. Let $\Lambda = \Sigma^{-1}$ be the corresponding concentration(precision) matrix. Given n independent and identically distributed observations (rows of $\mathbf{X}$), we try to find the undirected graph 'best' representing the conditional independence relationships among columns of $\mathbf{X}$.

For notational convenience, let us denote the jth column of $\mathbf{X}$ as $\mathbf{X}_j$. Thus, $\mathbf{X}_j = (X_{1j}, \ldots, X_{nj})^T$ and $\mathbf{X} = [\mathbf{X}_1, \ldots, \mathbf{X}_p]$. We further denote $\mathbf{p} = \{1, 2, \ldots, p\}$ and $\mathbf{X}_{\mathbf{p}\backslash\{t\}}$ is the matrix obtained after dropping the t-th column from $\mathbf{X}$.

The link between the pairwise Markov property and the entries of the inverse covariance matrix for a Gaussian random vector can formally be described as follows.

Lemma 3.1 *[Lauritzen [1996], page 129] Let* $\mathbf{p} = \{1, ..., p\}$. *Assume that* $\mathbf{X} \sim N_p(\mu, \Sigma)$, *where* Σ *is positive definite. Then it holds that*

$$\mathbf{X}_t \perp\!\!\!\perp \mathbf{X}_j | \mathbf{X}_{\mathbf{p}\setminus\{t,j\}} \Leftrightarrow \Lambda_{tj} = 0$$

There is a connection between pairwise Markov property and multiple regression as well. This partly follows from Lemma 3.1. In fact, it is known that for each $t \in p$, $\mathbf{X}_t$ can be represented as

$$\mathbf{X}_t = \sum_{j=1, j\neq t}^{p} \beta_{t,j} \mathbf{X}_j + \epsilon_t \tag{3.1.1}$$

where $\epsilon_t = (\epsilon_{t1}, ..., \epsilon_{tn})^T$ is independent of $\mathbf{X}_t$ and β_{tj} is the effect of node j on node t in the linear regression of all variables on $\mathbf{X}_t$.

It is well known Lauritzen [1996] that we can express β_{tj} and β_{jt} as

$$\beta_{tj} = \frac{-\Lambda_{tj}}{\Lambda_{jj}} = \rho_{tj.\mathbf{p}\setminus\{t,j\}} \sqrt{\frac{\Lambda_{tt}}{\Lambda_{jj}}}, \beta_{jt} = \frac{-\Lambda_{jt}}{\Lambda_{tt}} = \rho_{jt.\mathbf{p}\setminus\{t,j\}} \sqrt{\frac{\Lambda_{jj}}{\Lambda_{tt}}}$$

where $\rho_{tj.\mathbf{p}\setminus\{t,j\}}$ is the partial correlation between $\mathbf{X}_t$ and $\mathbf{X}_j$ given $\mathbf{X}_{\mathbf{p}\setminus\{t,j\}}$. In view of the two equations above, the following are equivalent. Note that $\beta_{tj} = 0$ if and only if $\beta_{jt} = 0$.

Theorem 3.1 *Let* $\mathbf{p} = \{1, ..., p\}$. *Assume that* $\mathbf{X} \sim N_p(\mu, \Sigma)$, *where* Σ *is positive definite. Then it holds that*

(1) $\mathbf{X}_t$ *and* $\mathbf{X}_j$ *is conditionally independent given* $\mathbf{X}_{\mathbf{p}\setminus(t,j)}$.

(2) $(t, j), (j, t) \notin E$.

(3) $\beta_{tj} = 0$ *and* $\beta_{tj} = 0$.

(4) $\Lambda_{tj} = 0$.

(5) $\rho_{tj.\mathbf{p}\setminus\{t,j\}} = 0$.

3.2 Model Selection for Undirected Graph

Numerous methods of model selection have been studied in literature. In method based on hypothesis testing, a huge number of test have to be done. This leads to two problems. First of all, it requires a huge computation time. Second, and more importantly, since a lot of hypothesis have to be tested, one quickly lands up in a multiple testing problem due to dependence among the test statistics maintaining a level might be difficult. Drton and Perlman [2004] use Sidek's inequality [Šidák, 1967] to test whether Fisher's z-transformed conditional correlations are equal to zero.

Penalization method, either directly penalizing the off-diagonal entries of the inverse covariance matrix or the regression coefficients in the equation 3.1.1, has been studied by several authors [Meinshausen and Bülmann, 2006, Yuan and Lin, 2007]. It is possible to penalize directly on $\rho_{tj.\mathbf{p}\setminus\{t,j\}}$ as well [Peng et al., 2009].

3.2.1 Direct penalization on Λ_{tj}

The likelihood function for multivariate Gaussian distribution depends on the precision matrix. Thus a natural approach would be to penalize the off diagonal entries of this precision matrix. In fact, Yuan and Lin [2007] proposed a procedure using a L_1 penalty on entries of the inverse covariance matrix. The procedure estimates Λ by the solution of the following constrained optimization problem,

$$\begin{aligned} &\min_{\mathbf{C}\in\mathcal{P}^+} -log|\mathbf{C}| + \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i^T\mathbf{C}\mathbf{x}_i \\ &\text{subject to } \sum_{t\neq j}|\mathbf{C}_{tj}| \leq t \end{aligned} \tag{3.2.1}$$

where $\mathcal{P}^+$ is the set of positive definite matrices and $\mathbf{C}_{tj}$ denotes (t, j) entry of $\mathbf{C}$. Equation (3.2.1) is the log-likelihood for Gaussian distribution. Originally, Yuan and Lin [2007] exploited the presence of logarithm in (3.2.1) and implemented the maxdet [Vandenberghe et al., 1998] procedure to find the estimate of Λ. This maxdet procedure ensures a global positive definite matrix as a minimizer for (3.2.1) but cannot handle

high dimensional data. Friedman et al. [2008] introduce the graphical LASSO algorithm which efficiently solve equation (3.2.1) when the number of variables is large. The glasso algorithm is efficient but due to its nonlinear nature, it is difficult to determine the solution path for all values of t.

3.2.2 Penalization on β_{tj}

Neighborhood Selection, introduced by Meinshausen and Bülmann [2006], uses LASSO to select the edges that is connected to each node. The neighborhood selection solves $\hat{\boldsymbol{\beta}}_t$ by taking

$$\hat{\boldsymbol{\beta}}_t = arg\min_{\beta_t} ||\mathbf{X}_t - \mathbf{X}_{-t}\beta_t||_2^2 + \lambda_t||\beta_t||_1$$

where

$$\boldsymbol{\beta}_t = (\beta_{t,1}, ..., \beta_{t,t-1}, \beta_{t,t+1}, ..., \beta_{t,p})^T,$$
$$\mathbf{X}_{-t} = (\mathbf{X}_1, ..., \mathbf{X}_{t-1}, \mathbf{X}_{t+1}, ..., \mathbf{X}_p).$$

Notice that the neighborhood selection does not ensure the symmetry of estimated *adjacency matrix* of the graph. That is to say, if node j is selected in the neighborhood of node t, there is no guarantee that the node t would be selected as a neighborhood of j.

In order to correct this problem, Meinshausen and Bülmann [2006] suggest *MB-OR* or *MB-AND* procedures. In the first one, an edge is selected if either $\beta_{tj} \neq 0$ or $\beta_{jt} \neq 0$. In the latter an edge is selected if both $\beta_{tj} \neq 0$ and $\beta_{jt} \neq 0$ hold. Consistency of MB-OR procedure with thresholding has been studied by Zhou et al. [2011].

3.2.3 Penalization on $\rho_{tj.\mathbf{P}\backslash\{t,j\}}$

A multiple regression based approach capable of selecting symmetric adjacency matrix was proposed by Peng et al. [2009]. Their method, called **SPACE**, is a joint sparse symmetric regression model estimation method. In particular, it involves solving the

problem,

$$\min_{\rho} \frac{1}{2} \sum_{t=1}^{p} \frac{1}{2} \left|\left| \mathbf{X}_t - \sum_{j=1, j\neq t}^{p} \rho_{tj.\mathbf{p}\setminus\{t,j\}} \sqrt{\frac{\Lambda_{tt}}{\Lambda_{jj}}} \mathbf{X}_j \right|\right|_2^2 + \lambda \sum_{1\leq t\leq j\leq p} |\rho_{tj.\mathbf{p}\setminus\{t,j\}}|.$$

The focus here is on the L_1 penalty [Tibshirani, 1996] of the partial correlations. Within the algorithm, SPACE alternates between estimating the partial correlation and residual variances. One of the major differences between neighborhood selection and SPACE is that the latter is symmetric and selects the neighborhoods of all the nodes together.

3.2.4 Symmetric LASSO and paired group LASSO

Friedman et al. [2010] propose two methods of estimating sparse graphical models. The first method, symmetric LASSO, involves symmetrizing the neighborhood selection approach, and is related to the SPACE method. Since $\mathbf{X}$ follows multivariate normal, the inverse covariance matrix captures the conditional distribution of each $\mathbf{X}_j$, given the other variables. Therefore, each β_{tj} can be re-parametrized in terms of the off-diagonal entries of Λ and the residual variance σ_{jj}^2. Using this property, symmetric LASSO estimates the off-diagonal entries of Λ and the residual variance by minimizing the negative log-product-likelihood for all the conditional distributions with the l_1 penalty of the entries in Λ. In particular, Friedman et al. [2010] propose to estimate the off-diagonal entries of Λ and the residual variance σ_{jj}^2 for $j = 1, .., p$ by taking the solution of the following optimization problem,

$$min_{\mathbf{C},\sigma_{11},\ldots,\sigma_{pp}} \frac{1}{n} \sum_{j=1}^{p} \left[nlog\sigma_{jj} + \frac{1}{\sigma_{jj}^2} ||\mathbf{X}_j + \mathbf{X}\mathbf{C}_j\sigma_{jj}||_2^2 \right] + \lambda \sum_{t<j} |\mathbf{C}_{tj}| \ \ s.t. \ \mathbf{C}_{tj} = \mathbf{C}_{jt}$$

where $\mathbf{C}$ is a p by p symmetric matrix with zeros on the diagonal and $\mathbf{C}_j$ is the jth column of $\mathbf{C}$. The above minimization problem holds because $\sigma_{jj}^2\beta_{tj} = \sigma_{tt}^2\beta_{jt}$. In here, $\mathbf{C}_{t,j} = 0$ implies that $\beta_{tj} = \beta_{jt} = 0$.

The second method is named paired group LASSO, an adaptation of the group LASSO

method to undirected graph selection. Paired group LASSO involves grouping β_{tj} and β_{jt} together. This ensures that any model selected from this method is symmetric. Similar to SPACE, paired group LASSO selects the neighborhood of all the nodes together.

3.3 Directed Acyclic Graphs

Recall that Directed acyclic graph only has directed edges and therefore if $(t, j) \in E$, $(j, t) \notin E$. Because of the directed edge, the Markov properties represented by a DAG are different those represented by an undirected graph. First of all, on a directed acyclic graph, two path-connected(Specified later) vertices can be unconditionally independent, which is not possible on an undirected graph.

Another property of a DAG is that it must be acyclic. Let $<_p$ and $>_p$ be binary relations that is defined as follows :

(1) $v <_p w$: For $(v, w) \in V \times V$, if there are $v_1, v_2, v_3, ..., v_k \in V$ such that $(v, v_1), (v_i, v_{i+1}), (v_{i+1}, w) \notin E$ and $(v_1, v), (v_{i+1}, v_i), (w, v_{i+1}) \in E$ for $i = 1, ..., k-1$.

(2) $v >_p w$: For $(v, w) \in V \times V$, if there are $v_1, v_2, v_3, ..., v_k \in V$ such that $(v_1, v), (v_{i+1}, v_i), (w, v_{i+1}) \notin E$ and $(v, v_1), (v_i, v_{i+1}), (v_{i+1}, w) \in E$ for $i = 1, ..., k-1$.

Also, for a pair of vertices, we say that $v \leq_p w$ if either $v = w$ or $v <_p w$.

A directed graph is acyclic if for any $v \in V$, $v \not>_p v$. That is, we cannot follow a sequence of directed arrow in one direction such that a node is cycled back to itself. We now introduce some preliminary notations for DAG.

3.3.1 Notations

For a vertex $v \in V$, define its parents, ancestors, children and descendant respectively by

$$pa(v) = \{x \in V : (x, v) \in E, (v, x) \notin E\},$$

$$an(v) = \{x \in V : v \leq_p x\},$$
$$ch(v) = \{x \in V : (x, v) \notin E, (v, x) \in E\},$$
$$de(v) = \{x \in V : v \geq_p x\}.$$

For any subset $V^* \subseteq V$, we define

$$pa(V^*) = \cup_{v \in V^*} pa(v).$$

The definition of $an(V^*)$, $ch(V^*)$ and $de(V^*)$ are similar to $pa(V^*)$.

The definition of path is required when establishing Markov properties in the next part. Before we can define a path on a directed acyclic graph, it is required to define its skeleton.

Definition 3.4 *For a DAG $G = (V, E)$, the skeleton of G is $G^* = (V, E^*)$ where E^* contains all (t, j) and (j, t) such that if $(t, j) \in E^*$.*

In other words, by replacing all the directed edges in a DAG with undirected edges, we get the skeleton.

Definition 3.5 *Let $G = (V, E)$ be an directed acyclic graph, and $G^* = (V, E^*)$ be the skeleton of G. For two distinct vertices a and c in V. A path π of length k is a set of k non-repeating vertices $v_1, ..., v_k$ such that $a = v_1$, $c = v_k$, and for every i from $1, ..., k-1$, $(v_i, v_{i+1}) \in E^*$ and $(v_{i+1}, v_i) \in E^*$.*

Notice that the path defined on the DAG does not need to follow the direction of edges. Therefore, any vertex on a path can be classified into two groups.

Definition 3.6 *Suppose that $G = (V, E)$ is a directed acyclic graph. A vertex v is a collider on a path π if there are two parents of v on π. Any vertex is a non-collider if it is not a collider.*

By our definition, a vertex is a non-collider on the path if either its an end-point of the path or it has at most one parent on the path.

$$a \rightarrow b \leftarrow c \rightarrow d \leftarrow e \rightarrow f$$

In the above graph, nodes a, c e and f are non-colliders while b and d are colliders on the path between a and f.

3.3.2 Markov Properties for directed acyclic graphs

Similar to undirected graphs, there are several list of Markov properties that can be described by DAG. A key concept in the directed global Markov property for DAG is defined on the moral graph, which we next define.

Definition 3.7 *Let $G = (V, E)$ be an directed acyclic graph. The moral graph G^m is obtained by placing an undirected edge between for every two nodes who have a common child, and then replacing all the directed edges with undirected edges.*

Suppose that $P(G)$ is a probability distribution defined on a DAG $G = (V, E)$, we say that P factorizes over G if its density f has the form [Lauritzen, 1996]

$$f(\mathbf{X}_1, ..., \mathbf{X}_p) = \prod_{i=1}^{p} f(\mathbf{X}_i | pa(\mathbf{X}_i)).$$

In fact, if P factorizes over G, it also obeys the global Markov property, which is defined as follows.

Definition 3.8 *P factorizes over G is equivalent to the directed global Markov property, which says that for disjoint sets $A \subset V$, $B \subset V$ and $S \subset V$, where S may be empty,*

$$A \perp\!\!\!\perp B | S$$

whenever A and B is separated by S in $(G_{an(A \cup B \cup S)})^m$, which is the moral graph of the smallest ancestral set containing $A \cup B \cup S$ (Lauritzen [1996], page 47).

It is known that the directed global Markov property is equivalent to the local directed Markov property [Lauritzen, 1996]. In particular, the local directed Markov property states that any variable v_i is conditional independent of its non-descendants, given its parents, i.e.

$$v_i \perp\!\!\!\perp V \backslash de(v_i) | pa(v_i).$$

An alternative to defining the directed global Markov property is using a path based d-connection criteria.

Definition 3.9 *For a DAG $G = (V, E)$, a path π between vertices a and c is said to be d-connecting given S (possibly empty) if*

(1) every non-collider on the path is not in S, and

(2) every collider on the path is in $an(S)$.

Definition 3.10 *For a DAG $G = (V, E)$, for disjoint sets A, B and S, where S may be empty, A and B are d-separated by S if for every $a \in A$ and $b \in B$, there is no path d-connecting a and c given S.*

Suppose $P(G)$ is a probability distribution defined on a DAG G, then we say that P satisfies the directed global Markov property if for any disjoint set A, B and S, $A \perp\!\!\!\perp B|S$ if and only if A is d-separated from B given S.

Also, it can be shown that the local directed Markov property is also equivalent to the ordered pairwise Markov property [Lauritzen, 1996]. The ordered pairwise Markov property requires the vertices of a DAG to be ordered. That is, the vertices in V are ordered and labeled as $\{1, ..., p\}$, then for $v_i, v_j \in V$

$$v_i \in pa(v_j) \Rightarrow v_i < v_j.$$

The ordered pairwise Markov property states that if the vertices are ordered, then for any $v_j \in V$

$$v_j \perp\!\!\!\perp pr(v_j) \backslash pa(v_j) | pa(v_j)$$

where $pr(v_i)$ is the predecessors of v_i, i.e. for any $v_i \in pr(v_j), v_i < v_j$.

3.3.3 Model selection for DAG

When the vertices in a DAG is ordered, we can retrieve the covariance matrix [Pourahmadi, 2000] for a Gaussian model by taking

$$\Sigma = B^{-1}D(B^T)^{-1} \tag{3.3.1}$$

where D and $B = (B_{tj})_{p\times p}$ are p by p matrices with

$$D = diag(\sigma_1^2, ..., \sigma_p^2),$$
$$B_{ij} = \begin{cases} 1 & \text{for } t = j. \\ -\beta_{tj} & \text{for } v_j \in pa(v_t). \\ 0 & otherwise. \end{cases}$$

B can be taken to be lower triangular because there is a natural ordering of vertices. In practice, when parameterizing a Gaussian DAG, people first specify the order among the vertices which can be read off from the direction of the arrows. Then, B, which is a lower triangular matrix, is specified, and Σ can be calculated from equation (3.3.1). The equivalence of the order pairwise property and directed global property implies that the parameterization will satisfy the conditional independence relationships specified by the d-separation criteria.

There are two approaches to model selection for DAG. The first approach is developed by Shojaie and Michailidis [2010]. They use a LASSO-based estimator to determine the parent of each node among the nodes. In here, we assume that the nodes are ordered. In particular, if $\mathbf{X}$ is Gaussian, the following set of structural regression equations hold [Pearl, 2000]. That is,

$$\mathbf{X}_t = \sum_{\mathbf{X}_j \in pa(\mathbf{X}_i)} \beta_{tj}\mathbf{X}_j + \epsilon_t \tag{3.3.2}$$

where $\beta_{t,j}$ is a regression coefficient that correspond to the edge $(j,t) \in E$ and ϵ_t is normally distributed with mean 0 and variance σ_t^2.

From equation (3.3.2), we can introduce a LASSO penalty and solve $\hat{\boldsymbol{\beta}}$ by taking the minimizer

$$\hat{\boldsymbol{\beta}}_t = argmin_{\beta_t}||X_t - \sum_{j=1}^{t-1} \beta_{t,j} X_j||_2^2 + \lambda_t \sum_{j=1}^{t-1} |\beta_{t,j}|.$$

Shojaie and Michailidis [2010] show that under certain assumptions, similar to Meinshausen and Bülmann [2006], these estimator is consistent.

The other approach to model selection for DAG is the PC-algorithm [Pearl [2000], Page 116]. The PC algorithm selects the edges of a DAG model by performing numerous pairwise conditional independence tests. One important assumption required for PC-algorithm is that the distribution is faithful to G and there must only be one graph such that the distribution is faithful. When we say that a distribution of faithful to a DAG, we say that all conditional independence relation for the distribution can be derived for d-separation.

There are other graphs like chain graph, mixed ancestor graph etc, but we don't discuss this.

CHAPTER 4

Edge Selection for Undirected Graph

4.1 Introduction

Graphical models provide an efficient way to represent and study complex statistical models. Each node of the graph usually represent a univariate random variable and the pattern of the edges represent conditional independence relationships between these random variables. Several graphical models using undirected, directed, mixed and bidirected edges have been studied in the literature. They are utilized in representing various combinations of conditional and unconditional independences among the nodes of the graph. In recent times, Graphical Markov models have been applied to many practical problems. Examples of such applications can be found in speech recognition, machine learning, environmental statistics and in recent times in genetic networks [Rodriguez-Concepcion and Boronat, 2002, Wille et al., 2004].

A major problem of interest in current statistics has been the selection of an appropriate graphical Markov model for given data set. Because of their construction and interpretation, different class of graphical models requires different techniques. Various

such class specific techniques for various graphical Markov model selection is known. In this chapter we focus on model selection techniques for Gaussian undirected graphs (UG).

Undirected graphs (UG) or concentration graphs [Dempster, 1972] is a large class of graphs Markov models which are represented by graphs with undirected edges. These graphs are useful in representing conditional independence relationships. Such graphs can be used to represent Markov random models in spatial statistics, models for social networks, models for genetic interactions etc. If it is reasonable to assume that the data follows a multivariate normal distribution, the absence of an edge between two nodes of the underlying UG implies that the corresponding off-diagonal element of the inverse covariance matrix or the precision matrix is zero. The converse is also true [Lauritzen, 1996].

Undirected graph selection or Covariance selection [Dempster, 1972] has received a great deal of attention from the researchers in recent times. From certain viewpoints model selection is akin to multiple hypothesis testing.

For Gaussian data selecting an UG is equivalent to testing if each pair of nodes are conditionally independent given all other nodes. Such conditional independence relationships correspond to zeros in the off-diagonal of the precision matrix for Gaussian data. Drton and Perlman [2004] use Sidek's inequality [Šidák, 1967] to test whether Fisher's z-transformed conditional correlations are equal to zero. They follow Holm's step down procedure [Holm, 1979] to select a UG from the p-values of the multiple tests. The testing based procedure described above requires the sample size to be much larger than the dimension of the data to be effective. Moreover, the number of hypotheses to test increases quadratically with the dimension. For similar procedures we refer to Drton and Perlman [2008].

A possible alternative to testing is to use a penalized likelihood based method with a penalty capable of shrinking some estimated parameters to zero. Several such penalty functions has been studied in literature. The most popular has been the L_1 penalty and the LASSO procedure introduced by Tibshirani [1996]. Others non-convex penalties like

SCAD [Fan and Li, 2001] have also been used recently.

Efficient methods to find the parameter estimates by maximizing these penalized likelihood are available. We refer to convex optimization algorithms described in Osborne et al. [2000], Efron et al. [2004], Friedman et al. [2007].

The likelihood function for multivariate Gaussian distribution depends on the precision matrix. Thus a natural approach would be to penalize the off diagonal entries of this precision matrix. Several authors have taken this route. Yuan and Lin [2007] impose L_1 penalty and use the *maxdet* algorithm [Vandenberghe et al., 1998] to find the penalized estimate of the precision matrix.

Their procedure require sample size to be larger than the dimension and cannot handle high dimensional data set. Friedman et al. [2008] implemented an efficient algorithm called Graphical Lasso, so that it can be applicable to high-dimensional data sets. Both methods can be slow due to their non-linear nature. Furthermore, finding the correct amount of shrinkage by cross-validation can often become troublesome.

Interestingly, multiple regression provides a convenient way to penalize off-diagonal entries in the precision matrix. It is well-known [Lauritzen, 1996] that if least square estimate of a regression parameter in a multiple regression problem is zero, the corresponding off-diagonal element in precision matrix of all the variables in the regression (including the response) is zero as well. This multiple regression based method does not use the Gaussian likelihood. Even though the connection of vanishing regression coefficients with absence of edges requires one to assume a Gaussian distribution. Meinshausen and Bülmann [2006] use this notion in a model selection method for undirected graphs. Each node is regressed on all the other nodes, with L_1 penalty imposed on the least square estimates of the regression parameters. It is seen that under certain conditions, using this method the true edge set can be consistently estimated. Multiple regression based method applies to high dimensional data. However, since this neighborhood selection method uses each node separately, the selected adjacency matrix can be asymmetric for finite sample sizes. Very often a post-selection symmetrization procedure is required.

A multiple regression based approach capable of selecting symmetric adjacency matrix was proposed by Peng et al. [2009]. Their method actually puts L_1 penalties on the off-diagonal elements of the precision matrix. The regression estimates are expressed in terms of the entries of the precision matrix and the usual least squared estimates are computed. This method, in each step, requires one to estimate the diagonal elements of the precision, which makes it quite slow.

Extending both the neighborhood selection and SPACE methods, Friedman et al. [2010] proposed symmetric LASSO and paired group LASSO. The symmetric LASSO is similar to SPACE, except that it puts L_1 penalties on the entries of the precision matrix instead of the partial correlation. The paired group LASSO uses the group LASSO method by Yuan and Lin [2006] and selects the neighborhoods of all the nodes together. Both methods ensures that any model selected would have a symmetric adjacency matrix.

We propose a new procedure called Edge Selection to identify the edges in an undirected graph. Motivated by Yuan and Lin [2006], it is done by grouping the variables together and applying group LARS on all possible edges together in the system. Edge Selection has two major advantages. The first is that the selected adjacency matrix is always symmetric. The second advantage is that the block-wise structure of the model specified in (4.3.1) ensures that this group LARS application is computationally efficient.

This chapter is setup as follows. In section 4.2, we introduce some basic notations and describe some of the methods that are already available for selecting undirected graphs. In section 4.3, we present the details of our proposed Edge selection algorithm. In section 4.4, we look at some properties of edge selection and also identify variables appropriate for cross-validation. In section 4.5, we look at the different types of cross validation that can be used with edge selection. Finally, the performance of Edge selection on both on simulated data and real data is discussed in section 4.6 and 4.7.

4.2 Background

4.2.1 Basic notations

Suppose $\mathbf{X}$ is a $n \times p$ data matrix, whose ith row $\mathbf{x}_i = (X_{i1}, \ldots, X_{ip})$ follows a multivariate normal with positive definite covariance matrix Σ, $1 \leq i \leq n$. Let $\Lambda = \Sigma^{-1}$ be our concentration matrix, where κ_{ij} represents the $(i,j)th$ entry of matrix κ. Given n independent and identically distributed observations (rows of $\mathbf{X}$), we try to find the undirected graph 'best' representing the conditional independence relationships among columns of $\mathbf{X}$.

For notational convenience, we denote the jth column of $\mathbf{X}$ as $\mathbf{X}_j$. Thus, $\mathbf{X}_j = (X_{1j}, \ldots, X_{nj})^T$ and $\mathbf{X} = [\mathbf{X}_1, \ldots, \mathbf{X}_p]$. We further denote $\mathbf{p} = \{1, 2, \ldots, p\}$ and $\mathbf{X}_{\mathbf{p}\setminus\{t\}}$ is the matrix obtained after dropping the tth column from $\mathbf{X}$. Let $\mathcal{E}$ be the underlying edge set and $\mathcal{E}_c = \{(t,j) : t \in \{1, ..., p-1]\}, j \in \{t+1, ..., p\}\}$ be set of all possible edges of the graph (i.e. the edge set of the complete graph). Also, let $\mathcal{E}^c = \mathcal{E}_c \setminus \mathcal{E}$.

As discussed in the last chapter, it is known that the pairwise Markov property [Lauritzen [1996], Meinshausen and Bülmann [2006]] $\mathbf{X}_t \perp\!\!\!\perp \mathbf{X}_j \mid \mathbf{X}_{\mathbf{p}\setminus\{t,j\}}$ holds if β_{tj} in the regression of $\mathbf{X}_t$ on the rest of the variables is zero. Moreover, this is equivalent to the *global Markov property* of the graph. Define β_{tj} as the effect of node j on node t, $j, t \in \mathbf{p}$, $j \neq t$. Note that, $\beta_{tj} = 0$ implies that $\beta_{jt} = 0$ as well.

4.3 Edge Selection

4.3.1 Setup

The proposed edge selection algorithm achieves two goals. First of all, it ensures symmetric selection at every step. Second, it selects the neighborhoods of all the nodes together.

Similar to Peng et al. [2009] we specify our model as a linear regression model where both the response and the auxiliary variables are derived from the data matrix $\mathbf{X}$. In

particular our model is specified as:

$$\mathbf{Y} = \mathbf{W}\boldsymbol{\beta} + \epsilon. \tag{4.3.1}$$

The response $\mathbf{Y}$ is obtained by column-wise vectoring $\mathbf{X}$. So $\mathbf{Y}_{np\times 1} = (\mathbf{X}_1, \ldots, \mathbf{X}_p)^T$. This is required for selecting the whole graph together.

The matrix of the covariates $\mathbf{W}$ is specifically constructed to allow symmetric selection. More specifically we define:

$$\mathbf{W}_{np\times p(p-1)} = [\mathbf{W}_{1,2}, \ldots, \mathbf{W}_{1,p}, \mathbf{W}_{2,3}, \ldots, \mathbf{W}_{2,p}, \mathbf{W}_{3,4}, \ldots, \mathbf{W}_{p-1,p}] \tag{4.3.2}$$

where for each $t = 1, 2, \ldots, p-1$ and $j = t+1, \ldots, p$, $W_{t,j}$ is a $np \times 2$ matrix constructed as:

$$\mathbf{W}_{t,j} = \begin{bmatrix} 0_{1\times n(t-1)} & \mathbf{X}_j^T & 0_{1\times n(t-j-1)} & 0_{1\times n} & 0_{1\times n(p-j-1)} \\ 0_{1\times n(t-1)} & 0_{1\times n} & 0_{1\times n(t-j-1)} & \mathbf{X}_t^T & 0_{1\times n(p-j-1)} \end{bmatrix}^T.$$

We center $\mathbf{Y}$ block-wise and standardize $\mathbf{W}$, such that $\mathbf{W}_{tj}^T\mathbf{W}_{tj} = I$.

In (4.3.1), $\epsilon_{np\times 1} = (\epsilon_{11}, ..., \epsilon_{1n}, ..., \epsilon_{pn})^T$ is the unknown vector of errors. By our definition, $\mathcal{E}_c = \{(t,j) : t \in \{1, ..., p-1]\}, j \in \{t+1, ..., p\}\}$. The parameter vector $\boldsymbol{\beta}$ can also be written as:

$$\boldsymbol{\beta}_{p(p-1)\times 1} = [\mathcal{B}_{1,2}, \ldots, \mathcal{B}_{1,p}, \mathcal{B}_{2,3}, \ldots, \mathcal{B}_{2,p}, \mathcal{B}_{3,4}, \ldots, \mathcal{B}_{p-1,p}]^T, \tag{4.3.3}$$

where for each $(t,j) \in \mathcal{E}_c$, $\mathcal{B}_{t,j} = [\beta_{tj}, \beta_{jt}]$. We define $\mathcal{B}_{t,j} = 0$ if and only if both $\beta_{tj} = 0$ and $\beta_{jt} = 0$ hold.

Notice that in the first column of matrix $\mathbf{W}_{t,j}$, $\mathbf{X}_j$ is located after $t-1$ blocks of zeros, where each block is of size n. Moreover, other then the t-th block, all other blocks in the first column of $\mathbf{W}_{t,j}$ are equals to zero. Therefore, $\mathbf{W}_{t,j}$ is constructed in such a way that β_{tj} is the regression coefficient for the effect of $\mathbf{X}_j$ on $\mathbf{X}_t$. The same applies to the second column.

The edge selection algorithm computes an estimate of $\boldsymbol{\beta}$ at every step. For any $\boldsymbol{\beta}$ we

define, $\boldsymbol{\mu}(\boldsymbol{\beta}) = \mathbf{W}\boldsymbol{\beta}$ and a $p(p-1) \times 1$ vector $\boldsymbol{c}(\boldsymbol{\beta})$ as:

$$\begin{aligned}\boldsymbol{c}(\boldsymbol{\beta}) &= \left[\mathbf{c}_{1,2}^T(\boldsymbol{\beta}), \ldots, \mathbf{c}_{1,p}^T(\boldsymbol{\beta}), \mathbf{c}_{2,3}^T(\boldsymbol{\beta}), \ldots, \mathbf{c}_{2,p}^T(\boldsymbol{\beta}), \mathbf{c}_{3,4}^T(\boldsymbol{\beta}) \ldots, \mathbf{c}_{p-1,p}^T(\boldsymbol{\beta})\right]^T \\ &= \mathbf{W}(\mathbf{Y} - \boldsymbol{\mu}(\boldsymbol{\beta})),\end{aligned}$$

where for $(t,j) \in \mathcal{E}_c$, $\mathbf{c}_{t,j}^T(\boldsymbol{\beta}) = \mathbf{W}_{t,j}^T(\mathbf{Y} - \mathbf{W}\boldsymbol{\beta}) = [c_{tj}(\boldsymbol{\beta}), c_{jt}(\boldsymbol{\beta})]$. Furthermore, let $C_{t,j}^2(\boldsymbol{\beta}) = \{c_{tj}(\boldsymbol{\beta})\}^2 + \{c_{jt}(\boldsymbol{\beta})\}^2$.

4.3.2 The Edge Selection Algorithm

The Edge selection algorithm performs in succession the operations described below:
Step (0) : **[Initialization.]** Initialize $k = 0$, $\hat{\boldsymbol{\beta}}^{(0)} = 0$, $\hat{\boldsymbol{\mu}}^{(0)} = \boldsymbol{\mu}\left(\hat{\boldsymbol{\beta}}^{(0)}\right) = 0$, $\hat{\mathbf{c}}^{(0)} = \boldsymbol{c}\left(\hat{\boldsymbol{\beta}}^{(0)}\right) = \mathbf{W}^T\mathbf{Y}$ and

$$\hat{E}_0 = \left\{(t_0, j_0) : C_{t_0,j_0}^2\left(\hat{\boldsymbol{\beta}}^{(0)}\right) = \max_{(t,j)}\left\{C_{t,j}^2\left(\hat{\boldsymbol{\beta}}^{(0)}\right)\right\}, (t,j) \in \mathcal{E}_c\right\}.$$

Further for any $(t_0, j_0) \in \hat{E}_0$, let $\mathcal{C}^2 = C_{t_0,j_0}^2\left(\hat{\boldsymbol{\beta}}^{(0)}\right)$.
Step (k, 1) : **[Target fixing.]** Assume the current values of $\hat{\boldsymbol{\beta}}^{(k)}$, $\hat{\boldsymbol{\mu}}^{(k)} = \boldsymbol{\mu}\left(\hat{\boldsymbol{\beta}}^{(k)}\right)$, $\hat{\mathbf{c}}^{(k)} = \boldsymbol{c}\left(\hat{\boldsymbol{\beta}}^{(k)}\right)$, $\mathcal{C}^2$ and $\hat{E}_k$ are available. Set $\hat{E} = \hat{E}_k$ and $\mathbf{W}_{\hat{E}} = [\mathbf{W}_{t,j}]_{(t,j)\in\hat{E}}$. We first compute $\bar{\boldsymbol{\beta}}_{\hat{E}}^{(k+1)}$ which solves

$$(\mathbf{W}_{\hat{E}}^T\mathbf{W}_{\hat{E}})\bar{\boldsymbol{\beta}}_{\hat{E}}^{(k+1)} = \mathbf{W}_{\hat{E}}^T\mathbf{Y}. \tag{4.3.4}$$

For a suitable permutation matrix $\mathbb{P}$ let $\bar{\boldsymbol{\beta}}^{(k+1)} = \mathbb{P} \times \left[\left(\bar{\boldsymbol{\beta}}_{\hat{E}}^{(k+1)}\right)^T, 0\right]^T$. Next calculate

$$\bar{\boldsymbol{\mu}}^{(k+1)} = \boldsymbol{\mu}\left(\bar{\boldsymbol{\beta}}^{(k+1)}\right) \text{ and } \boldsymbol{\theta} = \mathbf{W}^T(\bar{\boldsymbol{\mu}}^{(k+1)} - \hat{\boldsymbol{\mu}}^{(k)}). \tag{4.3.5}$$

Step (k, 2) : **[Edge inclusion.]** For each $(t', j') \notin \hat{E}$, compute

$$\gamma_{t'j'} = \min_{-,+}^{+} \left\{ \frac{-\left(\mathcal{C}^2 - P\right) \pm \sqrt{\left(\mathcal{C}^2 - P\right)^2 - \left(\mathcal{C}^2 - \Theta^2\right)\left(\mathcal{C}^2 - C'^2\right)}}{\mathcal{C}^2 - \Theta^2} \right\}, \tag{4.3.6}$$

where $C'^2 = \left\{\hat{c}^{(k)}_{t'j'}\right\}^2 + \left\{\hat{c}^{(k)}_{j't'}\right\}^2$, $\Theta^2 = \theta^2_{t'j'} + \theta^2_{j't'}$ and $P = \theta_{t'j'}\hat{c}^{(k)}_{j't'} + \theta_{j't'}\hat{c}^{(k)}_{j't'}$ and $\min_{-,+}^{+}$ is the minimum taken over the positive values of $\gamma_{t'j'}$ obtained from the two choices in $\pm$ sign.

Further suppose,

$$\gamma^{(k+1)} = \min_{(t',j')\notin\hat{E}} \left\{\gamma_{t'j'}\right\} \text{ and } (t_k, j_k) = arg\, min_{(t',j')\notin\hat{E}} \left\{\gamma_{t'j'}\right\}. \tag{4.3.7}$$

Step (k, 3) : **[Updating.]** Update $\hat{E}_{k+1} = \hat{E} \cup \{(t_k, j_k)\}$,

$$\hat{\boldsymbol{\beta}}^{(k+1)} = \hat{\boldsymbol{\beta}}^{(k)} + \gamma^{(k+1)}(\bar{\boldsymbol{\beta}}^{(k+1)} - \hat{\boldsymbol{\beta}}^{(k)}) \text{ and } \hat{\boldsymbol{\mu}}^{(k+1)} = \boldsymbol{\mu}\left(\hat{\boldsymbol{\beta}}^{(k+1)}\right). \tag{4.3.8}$$

Update $\hat{c}^{(k)}$ and $\mathcal{C}^2$ as

$$\hat{\mathbf{c}}^{(k+1)} = \boldsymbol{c}\left(\hat{\boldsymbol{\beta}}^{(k+1)}\right) = \mathbf{W}^T\left(\mathbf{Y} - \mathbf{W}\hat{\boldsymbol{\beta}}^{(k+1)}\right) \tag{4.3.9}$$

Step (k, 4) : **[Stopping rule.]** If $\mathcal{C}^2 = 0$, stop, otherwise, set $k \leftarrow k + 1$ return to Step (k, 1).

Equation (4.3.4) is always consistent, but $\mathbf{W}^T_{\hat{E}}\mathbf{W}_{\hat{E}}$ may be singular, especially when $n < p$. If $\mathbf{W}^T_{\hat{E}}\mathbf{W}_{\hat{E}}$ is singular, we take

$$\bar{\boldsymbol{\beta}}^{(k+1)}_{\hat{E}} = (\mathbf{W}^T_{\hat{E}}\mathbf{W}_{\hat{E}})^{+}\mathbf{W}^T_{\hat{E}}\mathbf{Y}$$

where $(\mathbf{W}^T_{\hat{E}}\mathbf{W}_{\hat{E}})^{+}$ is the Moore-Penrose inverse of $\mathbf{W}^T_{\hat{E}}\mathbf{W}_{\hat{E}}$. Obviously, if $\mathbf{W}^T_{\hat{E}}\mathbf{W}_{\hat{E}}$ is non-singular, then we use the usual $(\mathbf{W}^T_{\hat{E}}\mathbf{W}_{\hat{E}})^{-1}$. Our choice of Moore-Penrose inverse implies that $||\bar{\beta}^{(k+1)}||_2$ will be the minimum for any solution of equation (4.3.4). Furthermore, it ensures that $\mathcal{C}^2$ decreases all the way to 0 for $n < p$. We also observe that it is not

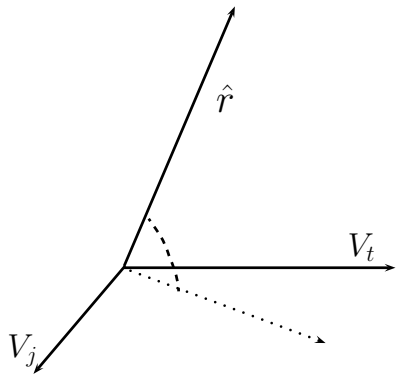


Figure 4.1 An illustration of an application of group LARS. Suppose we group vectors V_t and V_j, the angle between $\hat{r}$ and both V_t and V_j is the angle between $\hat{r}$ and its projection on V_t and V_j.

required to invert the whole matrix $\mathbf{W}_{\hat{E}}^T\mathbf{W}_{\hat{E}}$ as by definition, it is block-wise in structure. In fact, if there is only one edge added to the current step of the edge selection algorithm, it is required to invert at most only two sub matrices of maximum size $p-1$ by $p-1$.

4.4 Some properties of Edge Selection Algorithm

The Edge Selection Algorithm is a specific instance of the Group LARS algorithm [Yuan and Lin, 2006], which in turn is an extension to the LARS algorithm [Efron et al., 2004]. Starting from an empty model, in every step, LARS chooses the variables which have minimum angle with the then residual vector. In edge selection, in order to impose symmetry, we select two dimensional planes spanned by the columns of $\mathbf{W}_{t,j}$.

The angle between the residual vector $\hat{r}$ and its projection on the plane spanned by the columns of $\mathbf{W}_{t,j}$ is for this purpose. Since by our construction $\mathbf{W}_{t,j}^T\mathbf{W}_{t,j} = I$, it can be shown that the square of the cosine of this angle is proportional to $\hat{r}^T\mathbf{W}_{t,j}\mathbf{W}_{t,j}^T\hat{r}$. The algorithm allows more than one plane to be selected at each step. However, a plane once selected is never dropped.

4.4.1 Step-wise local properties of ES path

We now look at some properties of the edge selection path within each step of the algorithm. By construction, the edge selection satisfies the following property :

Property 1: Step-wise linearity of the shrunk parameter estimates: At any step k, $\hat{\boldsymbol{\beta}}^{(k+1)}$ is a linear combination of $\hat{\boldsymbol{\beta}}^{(k)}$ and $\bar{\boldsymbol{\beta}}^{(k+1)}$.

This property is evident from (4.3.8) in Step $(\mathsf{k}, 3)$. We shall show that, this linear combination is in fact a convex combination. To that effect, suppose at step k with $\hat{E}_k$ as the set of selected edges and for $\gamma \in \mathbb{R}$ define

$$\boldsymbol{\beta}^{(\gamma,k)} = \gamma\bar{\boldsymbol{\beta}}^{(k+1)} + (1-\gamma)\hat{\boldsymbol{\beta}}^{(k)}. \tag{4.4.1}$$

The following fundamental result follows from the construction:

Lemma 4.1 *For any* $(t,j) \in \hat{E}_k$, $\boldsymbol{\beta}_{t,j}^{(\gamma,k)} = \gamma\bar{\boldsymbol{\beta}}_{t,j}^{(k+1)} + (1-\gamma)\hat{\boldsymbol{\beta}}_{t,j}^{(k)}$ *implies* $\boldsymbol{c}_{t,j}\left(\boldsymbol{\beta}^{(\gamma,k)}\right) = (1-\gamma)\hat{c}_{t,j}^{(k)}$.

Proof: Let $\hat{E} = \hat{E}_k$. Using the definition of $\boldsymbol{c}_{tj}\left(\boldsymbol{\beta}^{(\gamma,k)}\right)$, we get

$$\begin{aligned}
\boldsymbol{c}_{tj}\left(\boldsymbol{\beta}^{(\gamma,k)}\right) &= \mathbf{W}_{t,j}^T(\mathbf{Y} - \mathbf{W}\boldsymbol{\beta}^{(\gamma,k)}) = \mathbf{W}_{t,j}^T\left\{\mathbf{Y} - \mathbf{W}_{\hat{E}}(\gamma\bar{\boldsymbol{\beta}}_{\hat{E}}^{(k+1)} + (1-\gamma)\hat{\boldsymbol{\beta}}_{\hat{E}}^{(k)})\right\} \\
&= \gamma\mathbf{W}_{t,j}(\mathbf{Y} - \mathbf{W}_{\hat{E}}\bar{\boldsymbol{\beta}}_{\hat{E}}^{(k+1)}) + (1-\gamma)\mathbf{W}_{t,j}^T(\mathbf{Y} - \mathbf{W}_{\hat{E}}\hat{\boldsymbol{\beta}}_{\hat{E}}^{(k)}) \\
&= (1-\gamma)\mathbf{W}_{t,j}^T(\mathbf{Y} - \mathbf{W}_{\hat{E}}\hat{\boldsymbol{\beta}}_{\hat{E}}^{(k)}) = (1-\gamma)\hat{\mathbf{c}}_{t,j}^{(k)}.
\end{aligned}$$

This completes Lemma 4.1. □

Lemma 4.1 implies that if $\boldsymbol{\beta}^{(\gamma,k)}$ is defined as in (4.4.1), then for any $(t,j) \in \hat{E}_k$, $C_{t,j}^2\left(\boldsymbol{\beta}^{(\gamma,k)}\right) = (1-\gamma)^2 C_{t,j}^2\left(\hat{\boldsymbol{\beta}}^{(k)}\right)$. This however is not true for any $(t',j') \notin \hat{E}_k$. That is, for $(t',j') \notin \hat{E}_k$, $C_{t',j'}^2\left(\hat{\boldsymbol{\beta}}^{(\gamma,k)}\right) \neq (1-\gamma)^2 C_{t',j'}^2\left(\hat{\boldsymbol{\beta}}^{(k)}\right)$.

The converse of Lemma 4.1 may not hold in a general setting. However, a partial converse for a special case, which is relevant to our Edge Selection algorithm holds.

Lemma 4.2 *Assume that for any* $(t', j') \notin \hat{E}_k$ *and* k, $\boldsymbol{\beta}^{(\gamma,k)}_{t'j'} = 0$ *and* $\bar{\boldsymbol{\beta}}^{(k+1)}_{t'j'} = 0$ *for all* $\gamma \in [0,1]$. *Suppose that* $\boldsymbol{c}_{t,j}\left(\boldsymbol{\beta}^{(\gamma,k)}\right) = (1-\gamma)\hat{\mathbf{c}}^{(k)}_{t,j}$. *Then* $(\mathbf{W}^T_{\hat{E}_k}\mathbf{W}_{\hat{E}_k})\boldsymbol{\beta}^{(\gamma,k)}_{\hat{E}_k} = (\mathbf{W}^T_{\hat{E}_k}\mathbf{W}_{\hat{E}_k})\left(\gamma\bar{\boldsymbol{\beta}}^{(k+1)}_{\hat{E}_k} + (1-\gamma)\hat{\boldsymbol{\beta}}^{(k)}_{\hat{E}_k}\right)$. *Furthermore for* $\gamma = 1$, *we have* $(\mathbf{W}^T_{\hat{E}_k}\mathbf{W}_{\hat{E}_k})\boldsymbol{\beta}^{(1,k)}_{\hat{E}_k} = \mathbf{W}^T_{\hat{E}_k}\mathbf{Y}$.

Proof: Let $\hat{E} = \hat{E}_k$. First of all note that by assumption $\boldsymbol{\beta}^{(\gamma,k)}_{j't'} = 0$ for all $(t', j') \notin \hat{E}$. Thus, $\mathbf{W}\boldsymbol{\beta}^{(\gamma,k)} = \mathbf{W}_{\hat{E}}\boldsymbol{\beta}^{(\gamma,k)}_{\hat{E}}$. Now, by definitions of $\boldsymbol{c}_{tj}\left(\boldsymbol{\beta}^{(\gamma,k)}\right)$ and $\hat{\mathbf{c}}^{(k)}_{t,j}$, we get

$$\boldsymbol{c}_{tj}\left(\boldsymbol{\beta}^{(\gamma,k)}\right) = \mathbf{W}^T_{t,j}(\mathbf{Y} - \mathbf{W}\boldsymbol{\beta}^{(\gamma,k)}) = \mathbf{W}^T_{t,j}(\mathbf{Y} - \mathbf{W}_{\hat{E}}\boldsymbol{\beta}^{(\gamma,k)}_{\hat{E}}) = \mathbf{W}^T_{t,j}\mathbf{Y} - \mathbf{W}^T_{t,j}\mathbf{W}_{\hat{E}}\boldsymbol{\beta}^{(\gamma,k)}_{\hat{E}}$$

and

$$(1-\gamma)\hat{\mathbf{c}}^{(k)}_{t,j} = (1-\gamma)\mathbf{W}^T_{t,j}\left[\mathbf{Y} - \mathbf{W}\hat{\boldsymbol{\beta}}\right] = (1-\gamma)\mathbf{W}^T_{t,j}\left[\mathbf{Y} - \mathbf{W}_{\hat{E}}\hat{\boldsymbol{\beta}}^{(k)}_{\hat{E}}\right].$$

Therefore, after some simplification, $\boldsymbol{c}_{tj}\left(\boldsymbol{\beta}^{(\gamma,k)}\right) = (1-\gamma)\hat{\mathbf{c}}^{(k)}_{t,j}$ implies that for $(t,j) \in \hat{E}$,

$$\mathbf{W}^T_{t,j}\mathbf{W}_{\hat{E}}\boldsymbol{\beta}^{(\gamma,k)}_{\hat{E}} = \gamma\mathbf{W}^T_{t,j}\mathbf{Y} - (1-\gamma)\mathbf{W}^T_{t,j}\mathbf{W}_{\hat{E}}\hat{\boldsymbol{\beta}}^{(k)}_{\hat{E}}. \tag{4.4.2}$$

Now by stacking (4.4.2) for to each $(t,j) \in \hat{E}$ and using the relation $\mathbf{W}^T_{\hat{E}}\mathbf{Y} = \mathbf{W}^T_{\hat{E}}\mathbf{W}_{\hat{E}}\bar{\boldsymbol{\beta}}^{(k+1)}_{\hat{E}}$ we get:

$$\begin{aligned}\mathbf{W}^T_{\hat{E}}\mathbf{W}_{\hat{E}}\boldsymbol{\beta}^{(\gamma,k)}_{\hat{E}} &= \gamma\mathbf{W}^T_{\hat{E}}\mathbf{Y} + (1-\gamma)\mathbf{W}^T_{\hat{E}}\mathbf{W}_{\hat{E}}\hat{\boldsymbol{\beta}}^{(k)}_{\hat{E}} = \gamma\mathbf{W}^T_{\hat{E}}\mathbf{W}_{\hat{E}}\bar{\boldsymbol{\beta}}^{(k+1)}_{\hat{E}} + (1-\gamma)\mathbf{W}^T_{\hat{E}}\mathbf{W}_{\hat{E}}\hat{\boldsymbol{\beta}}^{(k)}_{\hat{E}} \\ &= (\mathbf{W}^T_{\hat{E}}\mathbf{W}_{\hat{E}})\left(\gamma\bar{\boldsymbol{\beta}}^{(k+1)}_{\hat{E}} + (1-\gamma)\hat{\boldsymbol{\beta}}^{(k)}_{\hat{E}}\right).\end{aligned}$$

The second part follows trivially. □

Lemma 4.2 shows that if $\mathbf{W}^T_{\hat{E}_k}\mathbf{W}_{\hat{E}_k}$ is invertible, then $\boldsymbol{\beta}^{(\gamma,k)}_{\hat{E}_k} = \left(\gamma\bar{\boldsymbol{\beta}}^{(k+1)}_{\hat{E}_k} + (1-\gamma)\hat{\boldsymbol{\beta}}^{(k)}_{\hat{E}_k}\right)$ whenever $\boldsymbol{c}_{t,j}\left(\boldsymbol{\beta}^{(\gamma,k)}\right) = (1-\gamma)\hat{\mathbf{c}}^{(k)}_{t,j}$. Taking $\boldsymbol{\beta}^{(1,k)}_{\hat{E}_k} = (\mathbf{W}^T_{\hat{E}_k}\mathbf{W}_{\hat{E}_k})^+\mathbf{W}^T_{\hat{E}_k}\mathbf{Y}$, $\boldsymbol{c}_{t,j}\left(\boldsymbol{\beta}^{(\gamma,k)}\right) = (1-\gamma)\hat{c}^{(k)}_{t,j}$ implies that $\boldsymbol{\beta}^{(\gamma,k)}_{\hat{E}_k}$ must move towards $\bar{\boldsymbol{\beta}}^{(k+1)}_{\hat{E}}$.

Suppose $(t^\star, j^\star) \in \hat{E}_k$ and $\mathcal{C}^2 = C^2_{t^\star,j^\star}\left(\hat{\boldsymbol{\beta}}^{(k)}\right)$. For any such $(t', j') \notin \hat{E}$, the value of γ such that $C^2_{t',j'}\left(\hat{\boldsymbol{\beta}}^{(\gamma,k)}\right) = C^2_{t^\star,j^\star}\left(\hat{\boldsymbol{\beta}}^{(\gamma,k)}\right)$ can be found analytically.

Theorem 4.1 *With the notation in the algorithm and equation* (4.4.1)*, if* $(t_\star, j_\star) \in \hat{E}_k$*, then for any* $(t', j') \notin \hat{E}_k$*,* $C^2_{t',j'}\left(\boldsymbol{\beta}^{(\gamma,k)}\right) = C^2_{t^\star,j^\star}\left(\boldsymbol{\beta}^{(\gamma,k)}\right)$ *if*

$$\gamma = \left\{ \frac{-\left(\mathcal{C}^2 - P\right) \pm \sqrt{\left(\mathcal{C}^2 - P\right)^2 - \left(\mathcal{C}^2 - \Theta^2\right)\left(\mathcal{C}^2 - C'^2\right)}}{\mathcal{C}^2 - \Theta^2} \right\}. \tag{4.4.3}$$

Proof: Let $\boldsymbol{\mu}^{(\gamma,k)} = W\boldsymbol{\beta}^{(\gamma,k)}$, $r(\boldsymbol{\beta}^{(\gamma,k)}) = \mathbf{Y} - \mathbf{W}\boldsymbol{\beta}^{(\gamma,k)}$ and $r(\hat{\boldsymbol{\beta}}^{(k)}) = \mathbf{Y} - \mathbf{W}\hat{\boldsymbol{\beta}}^{(k)}$. By Lemma 4.1, $C^2_{t,j}(\boldsymbol{\beta}^{(\gamma,k)}) = (1-\gamma)^2\{(\hat{c}^{(k)}_{t,j})^2 + (\hat{c}^{(k)}_{t,j})^2\}$. So for $(t,j) \in \hat{E}$ and $(t', j') \notin \hat{E}$, solving $C^2_{t,j}(\boldsymbol{\beta}^{(\gamma,k)}) = C^2_{t',j'}(\boldsymbol{\beta}^{(\gamma,k)})$ is equivalent to solving

$$(1-\gamma)^2((\hat{c}^{(k)}_{t,j})^2 + (\hat{c}^{(k)}_{t,j})^2) = \left[r(\boldsymbol{\beta}^{(\gamma,k)})\right]^T \mathbf{W}_{t',j'}\mathbf{W}^T_{t',j'} r(\boldsymbol{\beta}^{(\gamma,k)}). \tag{4.4.4}$$

By definition of $r(\boldsymbol{\beta}^{(\gamma,k)})$,

$$\left[r(\boldsymbol{\beta}^{(\gamma,k)})\right]^T \mathbf{W}_{t',j'} = \left[\mathbf{Y} - \hat{\boldsymbol{\mu}}^{(k)} + \gamma(\bar{\boldsymbol{\mu}}^{(k+1)} - \hat{\boldsymbol{\mu}}^{(k)})\right]^T \mathbf{W}_{t',j'} = \left[\hat{c}^{(k)}_{t'j'} - \gamma\theta_{t'j'}, \hat{c}^{(k)}_{j't'} - \gamma\theta_{j't'}\right].$$

Thus the RHS of (4.4.4) equals:

$$\left[r(\boldsymbol{\beta}^{(\gamma,k)})\right]^T \mathbf{W}_{t',j'}\mathbf{W}^T_{t',j'} r(\boldsymbol{\beta}^{(\gamma,k)}) = (\hat{c}^{(k)}_{t'j'} - \gamma\theta_{t'j'})^2 + (\hat{c}^{(k)}_{j't'} - \gamma\theta_{j't'})^2. \tag{4.4.5}$$

Thus γ satisfies the equation

$$(1-\gamma)^2((\hat{c}^{(k)}_{tj})^2 + (\hat{c}^{(k)}_{tj})^2) = (\hat{c}^{(k)}_{t'j'} - \gamma\theta_{t'j'})^2 + (\hat{c}^{(k)}_{j't'} - \gamma\theta_{j't'})^2. \tag{4.4.6}$$

By simplifying (4.4.6) and using the definition of $\mathcal{C}^2$, Θ^2, P and C'^2 from the *edge inclusion* step of the edge selection algorithm (see (4.3.6)) we find that γ satisfies

$$\gamma^2\left(\mathcal{C}^2 - \Theta^2\right) - 2\gamma\left(\mathcal{C}^2 - P\right) + \left(\mathcal{C}^2 - C'^2\right) = 0. \tag{4.4.7}$$

Being a quadratic in γ, (4.4.7) readily yields solutions given by,

$$\gamma = \left\{ -\left(\mathcal{C}^2 - P\right) \pm \sqrt{\left(\mathcal{C}^2 - P\right)^2 - \left(\mathcal{C}^2 - \Theta^2\right)\left(\mathcal{C}^2 - C'^2\right)} \right\} / (\mathcal{C}^2 - \Theta^2). \tag{4.4.8}$$

□

Notice that, we get two possibly distinct values of γ in (4.4.3), obtained as solutions of a quadratic equation (see proof for details). Such analytic expressions are available, since in Edge selection we choose two dimensional planes spanned by the columns of $W_{t,j}$. This is not true for the the general group LARS, (compare Yuan and Lin [2006]) where γ would satisfy a polynomial equation of higher degree which could only be solved numerically.

In Step $(\mathsf{k}, 3)$ we update $\hat{E}_k$ by adding the edges in the set

$$\left\{(t_k, j_k) \;\; s.t. \; C^2_{t^\star, j^\star}\left(\hat{\boldsymbol{\beta}}^{(\gamma^{(k+1)},k)}\right) = C^2_{t_k,j_k}\left(\hat{\boldsymbol{\beta}}^{(\gamma^{(k+1)},k)}\right)\right\}. \tag{4.4.9}$$

Property 2: Path-wise maximality of $C^2_{t,j}(\boldsymbol{\beta})$ over the selected edge set: For any k, $(t,j) \in \hat{E}_k$ if and only if $C^2_{t,j}$ is maximal.

Property 2 follows from the construction and ensures two aspects of Edge selection at any step k. First of all, for any two edges (t,j) and $(t',j') \in \hat{E}_k$, $C^2_{t_k,j_k}\left(\boldsymbol{\beta}^{(\gamma,k)}\right) = C^2_{t'_k,j'_k}\left(\boldsymbol{\beta}^{(\gamma,k)}\right)$ for all γ. That is the value of $C^2_{t_k,j_k}$ remains constant over the $\hat{E}_k$ for all k. In the second place, if $(t,j) \in \hat{E}_k$ and for any $(t',j') \notin \hat{E}_k$, $C^2_{tj}\left(\hat{\boldsymbol{\beta}}^{(k)}\right) > C^2_{t',j'}\left(\hat{\boldsymbol{\beta}}^{(k)}\right)$.

Theorem 4.2 *For any $(t,j) \in \hat{E}_k$ and $(t',j') \notin \hat{E}_k$, at least one solution of $C^2_{t',j'}\left(\boldsymbol{\beta}^{(\gamma,k)}\right) = C^2_{t,j}\left(\boldsymbol{\beta}^{(\gamma,k)}\right)$ in γ is in $[0,1]$.*

Proof: In view of (4.4.6) we define

$$\mathcal{P}(\gamma) = (1-\gamma)^2\left\{(\hat{c}^{(k)}_{tj})^2 + (\hat{c}^{(k)}_{tj})^2\right\} - \left\{(\hat{c}^{(k)}_{t'j'} - \gamma\theta_{t'j'})^2 + (\hat{c}^{(k)}_{j't'} - \gamma\theta_{j't'})^2\right\}. \tag{4.4.10}$$

Clearly $\mathcal{P}$ is continuous in γ for all $\gamma \in \mathbb{R}$. Furthermore, from (4.4.10) and Property 2, $\mathcal{P}(0) > 0$, but $\mathcal{P}(1) \leq 0$. If $\mathcal{P}(1) = 0$, $\gamma = 1$ solves (4.4.6). If $\mathcal{P}(1) < 0$, by *Bolzano's theorem* [Apostol, 1997, Chapter 4.15], there is a $\gamma \in (0,1)$ such that $\mathcal{P}(\gamma) = 0$. □

Theorem 4.2 shows that out of two values of γ obtained from (4.4.3), at least one is in $[0,1]$. Thus for all k and for each $(t',j') \notin \hat{E}_k$, $\gamma_{t',j'} \in [0,1]$ in (4.3.6). So in (4.3.7), $\gamma^{(k+1)} \in [0,1]$ and $\hat{\boldsymbol{\beta}}^{(k+1)}$ is a convex combination of $\hat{\boldsymbol{\beta}}^{(k)}$ and $\bar{\boldsymbol{\beta}}^{(k+1)}$. Thus in Step $(\mathsf{k}, 2)$

the Edge selection selects the edges which correspond to the smallest root in $[0,1]$ of (4.4.9) for any $(t',j') \notin \hat{E}_k$.

4.4.2 Global properties of ES path

Even though, in each step of the algorithm, locally both $\hat{\boldsymbol{\beta}}^{(k)}$ and $\hat{\boldsymbol{c}}^{(k)}$ varies linearly, it is beneficial to know if such linearity holds with respect to some variable over the whole path. Such variables can be used for comparing two models on the selection path.

We define

$$\mathcal{M}_0(\boldsymbol{\beta}) = \max\left\{ \max_{(t,j),(j,t)} \{ \mid \mathbf{c}_{tj}(\boldsymbol{\beta}) \mid, \mid \mathbf{c}_{jt}(\boldsymbol{\beta}) \mid \} \;\; : \;\; (t,j) \in \hat{E}_0 \right\}. \tag{4.4.11}$$

By definition $\mathcal{M}_0$ is a function of the regression coefficient $\boldsymbol{\beta}$. For any $\boldsymbol{\beta}$, among all $(t,j) \in \hat{E}_0$, $\mathcal{M}_0(\boldsymbol{\beta}) = \mid c_{t_0 j_0}(\boldsymbol{\beta}) \mid$, where (t_0, j_0) is a maximal argument in (4.4.11). There are two possible ambiguities in this definition. First of all, the maximal argument in (4.4.11) may not be unique. Second, it is not clear if (t_0, j_0) changes with $\boldsymbol{\beta}$. The following result shows that such ambiguities can be easily resolved.

Lemma 4.3 *For any k and $\gamma \in [0,1]$ suppose $\hat{\boldsymbol{\beta}}^{(k)}$ is the current estimate of $\boldsymbol{\beta}$. Let $\boldsymbol{\beta}^{(\gamma,k)}$ be defined as in* (4.4.1)*. Then*

(1) $\mathcal{M}_0\left(\boldsymbol{\beta}^{(\gamma,k)}\right) = (1-\gamma)\mathcal{M}_0\left(\hat{\boldsymbol{\beta}}^{(k)}\right)$.

(2) $\mathcal{M}_0(\boldsymbol{\beta})$ *is completely determined at* Step 0 *of the algorithm.*

Proof: Let $\hat{E} = \hat{E}_k$. We have $\hat{\mathbf{c}}^{(k)} = \mathbf{W}'(\mathbf{Y} - \mathbf{W}\hat{\boldsymbol{\beta}}^{(k)})$ and $\hat{\mathbf{c}}^{(k)}_{\hat{E}} = \mathbf{W}'_{\hat{E}}(\mathbf{Y} - \mathbf{W}\hat{\boldsymbol{\beta}}^{(k)}) = \mathbf{W}'_{\hat{E}}(\mathbf{Y} - \mathbf{W}_{\hat{E}}\hat{\boldsymbol{\beta}}^{(k)}_{\hat{E}})$. In this case, $\hat{\mathbf{c}}^{(k)}_{\hat{E}}$ is obtained by taking the entries of $\hat{\mathbf{c}}^{(k)}$ that corresponds to $\hat{E}$. By Lemma 4.1, for any $(t,j) \in \hat{E}$, we have

$$\boldsymbol{c}_{t,j}\left(\boldsymbol{\beta}^{(\gamma,k)}\right) = (1-\gamma)\hat{c}^{(k)}_{t,j}$$

This implies that

$$\mathcal{M}_0\left(\boldsymbol{\beta}^{(\gamma,k)}\right) = \max\left\{ \max_{(t,j),(j,t)} \left\{ \mid c_{tj}\left(\boldsymbol{\beta}^{(\gamma,k)}\right) \mid, \mid c_{jt}\left(\boldsymbol{\beta}^{(\gamma,k)}\right) \mid \right\} \;\; : \;\; (t,j) \in \hat{E}_0 \right\}$$

$$= (1-\gamma)\max\left\{\max_{(t,j),(j,t)}\left\{|\ \hat{c}_{t,j}^{(k)}\ |,|\ \hat{c}_{j,t}^{(k)}\ |\right\}\ :\ (t,j)\in\hat{E}_0\right\}$$
$$= (1-\gamma)\mathcal{M}_0\left(\hat{\boldsymbol{\beta}}^{(k)}\right)$$

For (ii), suppose there are p_0 edges $(t_1,j_1),\ldots,(t_{p_0},j_{p_0})$ added at step 0 of the algorithm. WLOG, suppose if

$$\mathcal{M}_0\left(\hat{\boldsymbol{\beta}}^{(0)}\right) = |c_{t_m j_m}(\hat{\boldsymbol{\beta}}^{(0)})|$$

for a specific $m\in\{1,\ldots,p_0\}$. $\mathcal{M}_0(\boldsymbol{\beta})$ will be completely determined by Step 0 of the algorithm if

$$\mathcal{M}_0\left(\hat{\boldsymbol{\beta}}^{(k)}\right) = |c_{t_m j_m}(\hat{\boldsymbol{\beta}}^{(k)})|$$

By using Lemma 1, we have

$$\begin{aligned}\mathcal{M}_0\left(\hat{\boldsymbol{\beta}}^{(k)}\right) &= \max\left\{\max_{(t,j),(j,t)}\left\{|\ c_{tj}\left(\hat{\boldsymbol{\beta}}^{(k)}\right)|,|\ c_{jt}\left(\hat{\boldsymbol{\beta}}^{(k)}\right)|\right\}\ :\ (t,j)\in\hat{E}_0\right\}\\ &= \prod_{i=1}^{k}(1-\gamma^{(i)})\max\left\{\max_{(t,j),(j,t)}\left\{|\ c_{tj}\left(\hat{\boldsymbol{\beta}}^{(0)}\right)|,|\ c_{jt}\left(\hat{\boldsymbol{\beta}}^{(0)}\right)|\right\}\ :\ (t,j)\in\hat{E}_0\right\}\\ &= \prod_{i=1}^{k}(1-\gamma^{(i)})\mathcal{M}_0\left(\hat{\boldsymbol{\beta}}^{(0)}\right) = \prod_{i=1}^{k}(1-\gamma^{(i)})\mid c_{t_m j_m}(\hat{\boldsymbol{\beta}}^{(0)})\mid\\ &= |c_{t_k j_k}(\hat{\boldsymbol{\beta}}^{(k)})|\end{aligned}$$

with $\gamma^{(i)}$ from the ith step of the Edge Selection Algorithm from equation (4.3.7) □

Lemma 4.3 shows that, the pair (t_0,j_0) chosen at Step 0 of the algorithm, does not change with $\boldsymbol{\beta}$ over the path. Also, if at Step 0, there are two pairs (t_0,j_0) and $(t_{0'},j_{0'})$ such that $|\ c_{t_0,j_0}\ |=|\ c_{t_{0'},j_{0'}}\ |$, we can choose any one of them.

Lemma 4.3 also shows that in each step of the algorithm, the value of $\mathcal{M}_0$ decreases, that is, for each k, $\mathcal{M}_0\left(\hat{\boldsymbol{\beta}}^{(k+1)}\right)\le\mathcal{M}_0\left(\hat{\boldsymbol{\beta}}^{(k)}\right)$. Thus, along the path of the algorithm the function $\mathcal{M}_0$ shrinks to zero.

In the following theorem we show how functions like $\boldsymbol{\beta}^{(\gamma,k)}$, c etc. vary with $\mathcal{M}_0$.

Theorem 4.3 *Using the notation as described above at each step k of the algorithm, with $\hat{\boldsymbol{\beta}}^{(k)}$, $\bar{\boldsymbol{\beta}}^{(k+1)}$ fixed, the Edge selection algorithm satisfies:*

(1) $\boldsymbol{\beta}^{(\gamma,k)}$ *varies linearly with* $\mathcal{M}_0\left(\boldsymbol{\beta}^{(\gamma,k)}\right)$.

(2) For each $(t,j) \in \hat{E}_k$, *both* $\mid c_{tj}\left(\boldsymbol{\beta}^{(\gamma,k)}\right) \mid$ *and* $\mid c_{jt}\left(\boldsymbol{\beta}^{(\gamma,k)}\right) \mid$ *decreases linearly with decreasing* $\mathcal{M}_0\left(\boldsymbol{\beta}^{(\gamma,k)}\right)$ *Furthermore, for any* $l \geq k$, *the rate of reduction remains constant.*

(3) For each $(t,j) \in \hat{E}_k$, $C^2_{t,j}\left(\boldsymbol{\beta}^{(\gamma,k)}\right)$ *decreases when* $\mathcal{M}_0\left(\boldsymbol{\beta}^{(\gamma,k)}\right)$ *decreases.*

(4) At each step of the algorithm the residual sum of squares $S(\boldsymbol{\beta}) = ||\mathbf{Y} - \mathbf{W}\boldsymbol{\beta}||_2^2$ *decreases with decreasing* $\mathcal{M}_0$.

Proof: Note that from Lemma 4.3, we have

$$\gamma = \left(\mathcal{M}_0\left(\hat{\boldsymbol{\beta}}^{(k)}\right) - \mathcal{M}_0\left(\boldsymbol{\beta}^{(\gamma,k)}\right)\right) / \mathcal{M}_0\left(\hat{\boldsymbol{\beta}}^{(k)}\right) \tag{4.4.12}$$

From equation (4.4.12), and equation (4.3.8), we have

$$\boldsymbol{\beta}^{(\gamma,k)} = \hat{\boldsymbol{\beta}}^{(k)} + \gamma(\bar{\boldsymbol{\beta}}^{(k+1)} - \hat{\boldsymbol{\beta}}^{(k)}) = \bar{\boldsymbol{\beta}}^{(k+1)} - \frac{\mathcal{M}_0\left(\boldsymbol{\beta}^{(\gamma,k)}\right)}{\mathcal{M}_0\left(\hat{\boldsymbol{\beta}}^{(k)}\right)}(\bar{\boldsymbol{\beta}}^{(k+1)} - \hat{\boldsymbol{\beta}}^{(k)}) \tag{4.4.13}$$

Equation (4.4.13) varies linearly with $\mathcal{M}_0\left(\boldsymbol{\beta}^{(\gamma,k)}\right))$ as $\mathcal{M}_0\left(\hat{\boldsymbol{\beta}}^{(k)}\right)$, $\bar{\boldsymbol{\beta}}^{(k)}_{\hat{E}}$ and $\hat{\boldsymbol{\beta}}^{(k)}_{\hat{E}}$ are all fixed.

For (ii), Lemma 4.3 and part 1 of Theorem 4.3 implies that $\boldsymbol{c}_{t,j}\left(\boldsymbol{\beta}^{(\gamma,k)}\right)$ can be represented as

$$|\boldsymbol{c}_{tj}\left(\boldsymbol{\beta}^{(\gamma,k)}\right)| = \frac{\mathcal{M}_0\left(\boldsymbol{\beta}^{(\gamma,k)}\right)}{\mathcal{M}_0\left(\hat{\boldsymbol{\beta}}^{(k)}\right)}|\hat{\mathbf{c}}^{(k)}_{tj}|.$$

Thus, $|\boldsymbol{c}_{tj}\left(\boldsymbol{\beta}^{(\gamma,k)}\right)|$ decreases linearly with $\mathcal{M}_0\left(\boldsymbol{\beta}^{(\gamma,k)}\right)$. Also, suppose edge (t,j) is added at step k, then for any l such that $l \geq k$,

$$\frac{|\boldsymbol{c}_{tj}\left(\boldsymbol{\beta}^{(\gamma,l)}\right)|}{\mathcal{M}_0\left(\boldsymbol{\beta}^{(\gamma,l)}\right)} = \frac{|\boldsymbol{c}_{tj}\left(\hat{\boldsymbol{\beta}}^{(l)}\right)|}{\mathcal{M}_0\left(\hat{\boldsymbol{\beta}}^{(l)}\right)} = \frac{|\boldsymbol{c}_{tj}\left(\hat{\boldsymbol{\beta}}^{(l)}\right)|}{|\boldsymbol{c}_{t_0j_0}\left(\hat{\boldsymbol{\beta}}^{(l)}\right)|} = \frac{\prod_{i=l}^{k}(1-\gamma^{(i)})|\boldsymbol{c}_{tj}\left(\hat{\boldsymbol{\beta}}^{(k)}\right)|}{\prod_{i=l}^{k}(1-\gamma^{(i)})|\boldsymbol{c}_{t_0j_0}\left(\hat{\boldsymbol{\beta}}^{(k)}\right)|} = \frac{|\boldsymbol{c}_{tj}\left(\hat{\boldsymbol{\beta}}^{(k)}\right)|}{|\boldsymbol{c}_{t_0j_0}\left(\hat{\boldsymbol{\beta}}^{(k)}\right)|}.$$

Thus, the rate of reduction remains constant for any $l \geq k$.

For (iii), from part 2 of Theorem 4.3, we know that $|\hat{\mathbf{c}}_{tj}|$ and $|\hat{\mathbf{c}}_{jt}|$ decreases linearly

with respect to $\mathcal{M}_0\left(\boldsymbol{\beta}^{(\gamma,k)}\right)$, therefore $\hat{\mathbf{c}}_{tj}^2 + \hat{\mathbf{c}}_{jt}^2$ decreases when $\mathcal{M}_0\left(\boldsymbol{\beta}^{(\gamma,k)}\right)$ decreases.

For (iv), using Jensen's inequality and noting that $||\mathbf{Y} - \mathbf{W}\bar{\boldsymbol{\beta}}^{(k)}||_2^2 \leq ||\mathbf{Y} - \mathbf{W}\hat{\boldsymbol{\beta}}^{(k)}||_2^2$, we have

$$||\mathbf{Y} - \mathbf{W}\boldsymbol{\beta}^{(\gamma,k)}||_2^2 \leq (1-\gamma)||\mathbf{Y} - \mathbf{W}\hat{\boldsymbol{\beta}}^{(k)}||_2^2 + \gamma||\mathbf{Y} - \mathbf{W}\bar{\boldsymbol{\beta}}^{(k)}||_2^2 \leq ||\mathbf{Y} - \mathbf{W}\hat{\boldsymbol{\beta}}^{(k)}||_2^2$$

Therefore $S(\boldsymbol{\beta}^{(\gamma,k)}) \leq S(\hat{\boldsymbol{\beta}}^{(k)})$. This completes Theorem 4.3. □

The results in Theorem 4.3 are illustrated in Figure 4.2 above, where we consider a Gaussian first order autoregressive model on 3 nodes and sample size 10. Each diagonal element of the precision matrix Λ are taken to be equal to 1 and $\Lambda_{2,1} = \Lambda_{1,2} = \Lambda_{2,3} = \Lambda_{3,2} = 0.1$. From the model $\Lambda_{3,1} = \Lambda_{1,3} = 0$.

Figures 4.2(a), 4.2(b) and 4.2(c) respectively shows the path for the regression coefficients, vector $\mathbf{c}$ and C^2. The abscissa for each plot is $\mathcal{M}_0$ who are to be viewed from right to left. There are at most three edges. The edge $(1,2)$, $(1,3)$ and $(2,3)$ are respectively selected at $\mathcal{M}_0$ equal to 1.726, 1.284 and 0.819.

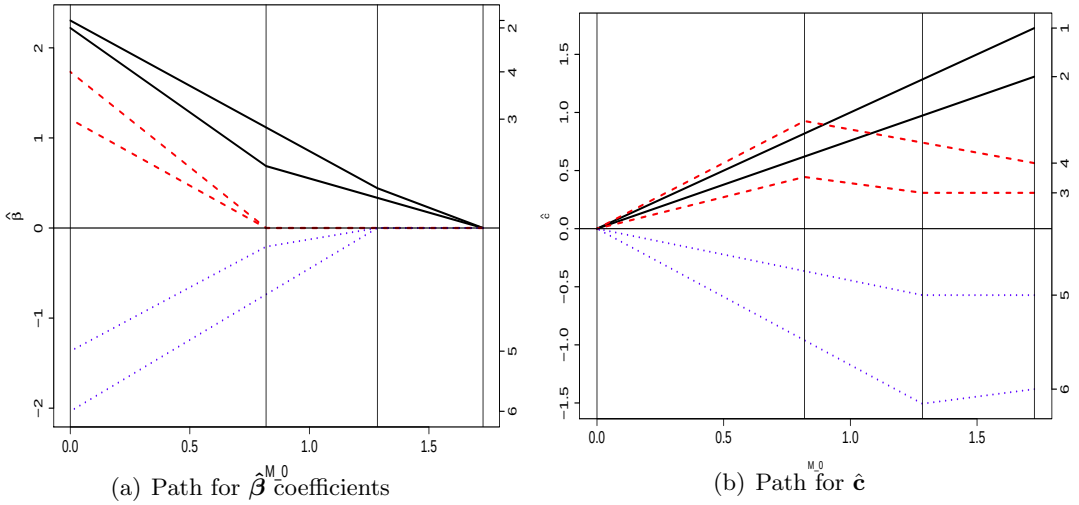
From Theorem 4.3 we get that for each k and for any $\gamma \in [0,1]$, $\boldsymbol{\beta}^{(\gamma,k)}$ is linear with respect to $\mathcal{M}_0\left(\boldsymbol{\beta}^{(\gamma,k)}\right)$. However, the slope and the intercept of this relationship depends on k. Thus, the estimated regression coefficients are not linear with respect to $\mathcal{M}_0$. They are only piece-wise linear (see Figure 4.2(a)). From the same figure it seems that the absolute value of parameter estimates increase along the path. This is not true in general. From the proof we can however can find a lower bound for $\mid \boldsymbol{\beta}^{(\gamma,k)} \mid$ as:

$$\left|\boldsymbol{\beta}^{(\gamma,k)}\right| \;\geq\; \left|\bar{\boldsymbol{\beta}}^{(k+1)}\right| \;-\; \left(\frac{\mathcal{M}_0\left(\boldsymbol{\beta}^{(\gamma,k)}\right)}{\mathcal{M}_0\left(\hat{\boldsymbol{\beta}}^{(k)}\right)}\right)\left|\bar{\boldsymbol{\beta}}^{(k+1)} - \hat{\boldsymbol{\beta}}^{(k)}\right|. \tag{4.4.14}$$

Notice that, this lower bound (ie. the right hand side of (4.4.14)) increases as $\mathcal{M}_0\left(\boldsymbol{\beta}^{(\gamma,k)}\right)$ decreases. However, it does not imply that $\mid \boldsymbol{\beta}^{(\gamma,k)} \mid$ will increase along the path.

After an edge (t,j) is selected in the edge set, it is never dropped and the vector $\mathbf{c}_{t,j}$ shrinks linearly to zero with $\mathcal{M}_0$ for the rest of the path (see Figure 4.2(b)). Naturally $C_{t,j}^2$ decreases to zero along $\mathcal{M}_0$, in a possibly quadratic fashion (see Figure 4.2(c)).

Figure 4.2 illustrates few other things. First of all, in Figures 4.2(b) and 4.2(c) we



(a) Path for $\hat{\boldsymbol{\beta}}$ coefficients

(b) Path for $\hat{\mathbf{c}}$

(c) Path for C^2

Figure 4.2 Edge Selection path of a first order autoregressive model with three nodes and sample size 10, with respect to $\mathcal{M}_0$. The Edge selection algorithm moves from right to left.

see that for edges not in the current edge set, $\mid c_{tj} \mid$, $\mid c_{jt} \mid$ and C^2_{tj} actually increases along the path. However, once they are selected these three functions start to decrease. Moreover, $C^2_{t,j}$ is constant for all selected edges (t, j).

Another curious fact seen in Figure 4.2(b) is that for $k > 0$, it is possible that for

some $(t, j) \in A_k$ one of $\mid c_{tj} \mid$ and $\mid c_{jt} \mid$ is larger than $\mathcal{M}_0\left(\hat{\boldsymbol{\beta}}^{(k)}\right)$. It is not possible for both $\mid c_{tj} \mid$ and $\mid c_{jt} \mid$ to be larger than $\mathcal{M}_0\left(\hat{\boldsymbol{\beta}}^{(k)}\right)$ as it would violate property 2 in Section 4.4.1.

In LASSO, the selection parameter is usually λ, which controls the amount of regularization applied to the coefficient estimates. A large value of λ can completely shrink some of the coefficient estimates to zero, while setting $\lambda = 0$ converts the LASSO problem to an Ordinary Least Squares problem. In LARS, the sum of the absolute of the coefficient estimates is usually used as a selection parameter.

There are similarities between $\mathcal{M}_0(\boldsymbol{\beta})$ and the λ parameter in LASSO. The latter can be viewed as the maximum correlation between all auxiliary variable with the current residual. In particular, statements similar to Lemma 3 and part (a) and (d) of Theorem 4.3 would hold as well.

In the next section, we discuss different methods for choosing an appropriate value of $\mathcal{M}_0(\boldsymbol{\beta})$ on the path.

4.5 Methods for choosing a model from the Edge selection path

4.5.1 Notations

The Edge Selection algorithm discussed above, traces the whole path and at each step $k = 0, 1, \ldots, p(p-1)/2$, produces $\hat{\boldsymbol{\beta}}^{(k)}$, $\mathcal{M}_0\left(\hat{\boldsymbol{\beta}}^{(k)}\right)$, $\boldsymbol{c}\left(\hat{\boldsymbol{\beta}}^{(k)}\right)$, $\hat{E}_k$ and $\bar{\boldsymbol{\beta}}^{(k+1)}$. We still need to choose a specific model on the path.

Note that $\mathcal{M}_0$ decreases linearly to zero along the path. Theorem 4.3 shows that at each step $\boldsymbol{\beta}^{(\gamma,k)}$ varies linearly with $\mathcal{M}_0\left(\boldsymbol{\beta}^{(\gamma,k)}\right)$. Thus we use $\mathcal{M}_0$ to choose our model on the path. Our choice of $\mathcal{M}_0$ is in the same spirit to LASSO where the tuning parameter is related to the highest correlation between the residual vector and the covariates.

For any given m, we compute,

$$k_m = \underset{k\geq 0}{\operatorname{argmin}}^{+} \left(\mathcal{M}_0 \left(\hat{\boldsymbol{\beta}}^{(k)} \right) - m \right) \text{ and } \gamma_m = \frac{\mathcal{M}_0 \left(\hat{\boldsymbol{\beta}}^{(k_m)} \right) - m}{\mathcal{M}_0 \left(\hat{\boldsymbol{\beta}}^{(k_m)} \right)}, \tag{4.5.1}$$

4.5.2 Multifold cross validation based methods

For multifold cross validation the rows of data matrix $\mathbf{X}$ are first randomly split into B different sets, $\mathbf{X}_1^\star, \ldots, \mathbf{X}_B^\star$ of sizes $n_1, \ldots, n_B$. Suppose $\mathbf{Y}_1^\star, \ldots, \mathbf{Y}_B^\star$ and matrices $\mathbf{W}_1^\star, \ldots, \mathbf{W}_B^\star$ are the corresponding response vectors and the matrix of covariates as described in Section 4.3. For any $b = 1, 2, \ldots, B$, let $\mathbf{X}_{-b}^\star$ be the data matrix obtained after removing $\mathbf{X}_b^\star$, with $\mathbf{Y}_{-b}^\star$ and matrix $\mathbf{W}_{-b}^\star$ denoting respectively the corresponding vector of responses and the matrix of covariates. Let $\hat{\boldsymbol{\beta}}_{-b}^{(k)}$ be the coefficient estimate obtained from equation (4.3.8), based on $\mathbf{Y}_{-b}^\star$ and matrix $\mathbf{W}_{-b}^\star$. For each $\mathcal{M}_0 = m$ with a function $\mathcal{L}$ depending on the data and $\boldsymbol{\beta}$ chosen beforehand, we define:

$$R_b(m) = \mathcal{L}\left(\mathcal{X}_b^\star, \boldsymbol{\beta}_{-b,m}^\star\right) \tag{4.5.2}$$

$$\bar{R}(m) = \sum_{b=1}^{B} R_b(m)/B \tag{4.5.3}$$

$$se(m) = \left[\sum_{b=1}^{B} \{R_b(m) - \bar{R}(m)\}^2 / B(B-1) \right]^{1/2}. \tag{4.5.4}$$

The following three cross validation methods can be used with different choices of $\mathcal{L}$ and $\boldsymbol{\beta}^\star$.

ES.CV$_{\text{e}}$ **:** For $\mathcal{M}_0 = m$, let k_m and γ_m be defined as in (4.5.1). Take

$$\boldsymbol{\beta}_{-b,m}^\star = \boldsymbol{\beta}_{-b}^{(\gamma_m, k_m)} \text{ and } \mathcal{L}\left(\mathbf{X}_{-b}^\star, \boldsymbol{\beta}_{-b,m}^\star\right) = || \mathbf{Y}_b^\star - \mathbf{W}_b^\star \boldsymbol{\beta}_{-b,m}^\star ||_2^2 / n_b. \tag{4.5.5}$$

With these choices of $\boldsymbol{\beta}_{-b,m}^\star$ and $\mathcal{L}$, suppose $m^\star$ minimises $\bar{R}(m)$ in (4.5.3). By following [Breiman et al., 1984] we select the model corresponding to the largest m such that $\bar{R}(m) \leq \bar{R}(m^\star) + e.se(m^\star)$. The constant e, usually assumed to be 1

or 2, controls the sparsity of the selected model. Our choice of m is clearly biased towards more sparse graphs and will reduce the number of edges selected. However, it is known that this method works well for shrunk estimates, which are similar to our group LARS estimates.

ES.CV$_{\min}$ **:** Special case of ES.CV$_{\mathrm{e}}$, where $e = 0$.

ES.OLS **:** Here we take $\boldsymbol{\beta}^{\star}_{-b,m} = \bar{\boldsymbol{\beta}}^{(k_m+1)}_{-b}$ and use the same $\mathcal{L}$ as in (4.5.5). In this case we choose the model corresponding to m which minimizes $\bar{R}(m)$ in (4.5.3). It is trivial to note that, since we use the OLS estimator of $\bar{\boldsymbol{\beta}}^{(k_m+1)}_{-b}$ in $\mathcal{L}$, for a fixed b, $R_b(m)$ is piece-wise constant in m.

ES.IPF **:** The fourth method uses Iterative Proportional fitting(IPF) [Whittaker, 1990], [Speed and Kiiveri, 1986] to estimate a covariance matrix preserving the given structure of the undirected graph. For $\mathcal{M}_0 = m$, let $\hat{E}_{k_m}$ be the set of selected edges obtained from the Edge selection algorithm. We use $\mathcal{X}^{\star}_{-b}$ and IPF algorithm to estimate $\hat{\Sigma}_{k_m}$, the covariance matrix of the UG corresponding to $\hat{E}_{k_m}$.

Now let S_b be the sample covariance matrix of $\mathcal{X}^{\star}_{b}$. We use the Kullback-Leibler divergence between $\hat{\Sigma}_{k_m}$ and S_b as $R_b\,(m)$. That is,

$$R_b(m) = -log(det(\hat{\Sigma}_{k_m})) + tr(\hat{\Sigma}^{-1}_{k_m} S_b).$$

The function $\bar{R}(m)$ is defined as in (4.5.3). We choose $\hat{E}_{k_m}$ corresponding to m which minimizes $\bar{R}(m)$. For IPF, one assumption needed is that $n > p$.

4.6 Simulation Study

4.6.1 Measures of comparisons and models

In this section we compare the proposed Edge selection algorithm (ES) with MB-AND, MB-OR and the SPACE algorithm discussed in [Peng et al., 2009]. In the SPACE method, three different type of weights were used, namely SPACE (or SPACE.NULL),

SPACE.SW and SPACE.DEW. SPACE uses equal weights, while SPACE.SW and SPACE.DEW use weights equal to the residual variance and weights proportional to the estimated degree of each nodes respectively.

We first compare the number of true edges selected before a fixed proportion of possible false edges selected by various methods. This is done without any cross-validation. Then, we consider the performance of the proposed Edge Selection algorithm with the multi-fold cross validation techniques described in Section 4.5.2. In our simulations, we consider three different number of nodes, $p = 10$, $p = 15$ and $p = 30$ with varying sample sizes.

Let $\hat{E}$ be the estimate edge set, and $\hat{E}^c = \mathcal{E}_c \backslash \hat{E}$. For a set A, let $\#(A)$ denote the number of elements in A. Measure of comparisons used are True Positive $TP = \#(\hat{E} \cap \mathcal{E})$, False Positive $FP = \#(\hat{E} \cap \mathcal{E}^c)$, True Negatives $TN = \#(\hat{E}^c \cap \mathcal{E}^c)$, False Negatives $FN = \#(\hat{E}^c \cap \mathcal{E})$ and the Matthews correlation coefficient (MC) [Shojaie and Michailidis, 2010] as

$$MC = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}.$$

For the purpose of this simulation study, we consider nine models described below. Each model is parametrized by the precision matrix Λ. Note that, by definition, absence of an edge between two nodes on an UG implies the corresponding entry in Λ is equal to zero. The specific values in the undirected graphs below are chosen to ensure a positive definite Λ. The model considered are as follows :

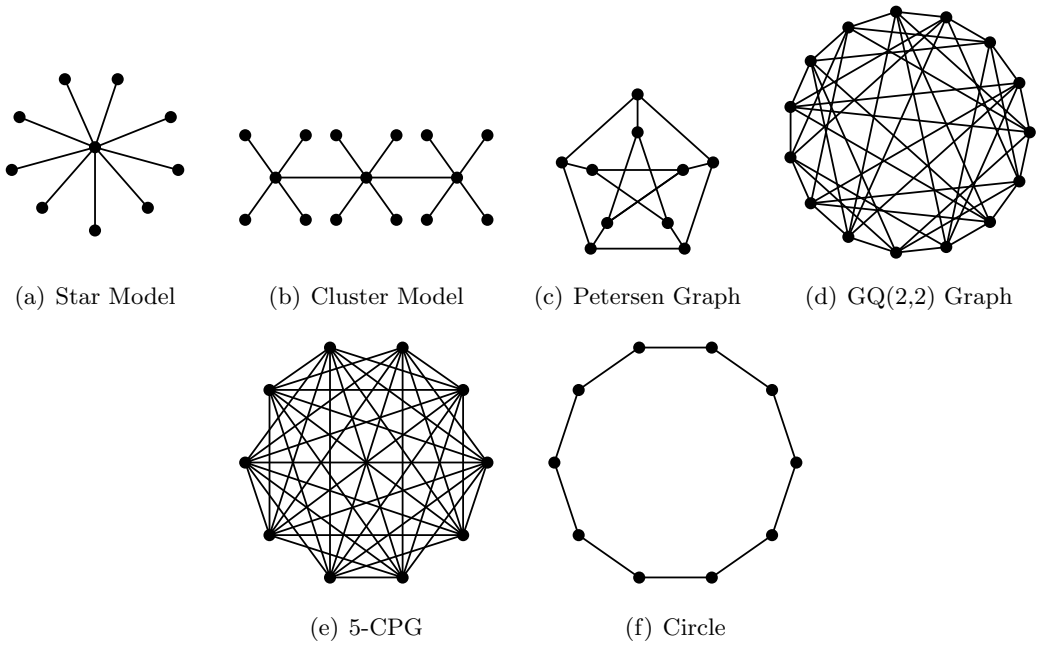
(1) AR(1) Model with $\Lambda_{i,i} = 1$ and $\Lambda_{i,i-1} = \Lambda_{i-1,i} = 0.5$, $i = 1, ..., p$.

(2) AR(2) Model with $\Lambda_{i,i} = 1$, $\Lambda_{i,i-1} = \Lambda_{i-1,i} = 0.5$ and $\Lambda_{i,i-2} = \Lambda_{i-2,i} = 0.25$, $i = 1, ..., p$.

(3) AR(4) Model with $\Lambda_{i,i} = 1$, $\Lambda_{i,i-1} = \Lambda_{i-1,i} = 0.4$, $\Lambda_{i,i-2} = \Lambda_{i-2,i} = \Lambda_{i,i-3} = \Lambda_{i-3,i} = 0.2$ and $\Lambda_{i,i-4} = \Lambda_{i-4,i} = 0.1$, $i = 1, ..., p$.

(4) Star Model with every node connected to the first node (see Figure 4.3(a)), we assign $\Lambda_{i,i} = 1$, $\Lambda_{1,j} = \Lambda_{j,1} = 0.2$ for all $j \neq 1$.

(5) A cluster Model consisting of a collection of star models with 5 nodes. The



(a) Star Model (b) Cluster Model (c) Petersen Graph (d) GQ(2,2) Graph

(e) 5-CPG (f) Circle

Figure 4.3

hub of the stars are arranged in an AR(1) formation (see Figure 4.3(b)). For parameterization, we assume $\Lambda_{t,t+5} = \Lambda_{t+5,t} = 0.5$, $\Lambda_{i,i} = 1$ and $\Lambda_{t,j} = \Lambda_{j,t} = 0.2$ for all $t = 5m + 1$ for all positive integer m, $t + 1 \leq j \leq t + 4$.

(6) Petersen Graph (Holton and Sheehan [1993], see Figure 4.3(c)) with $\Lambda_{t,j} = 0.3$ for all t and j if t is adjacent to j,

(7) Generalized Quadrangle Graph GQ(2,2) [Payne, 1973], the smallest non-trivial generalized quadrangle graph (see Figure 4.3(d)). We again assume $\Lambda_{t,j} = 0.3$ for all adjacent vertices t and j.

(8) q-Cocktail Party graph (q-CPG), where two sets each containing q nodes paired against each other. All nodes are connected except the paired ones (see Figure 4.3(e)). Here, we assume $\Lambda_{2i+1,j} = 0.3$, $\Lambda_{2i+2,2i+j} = 0.3$, $i = 0, ..., p - 1$ and $j = 2i + 3, 2i + 4...p - 1, p$, where $p = 2q$.

(9) Circle model (see Figure 4.3(f)) with $\Lambda_{i,i} = 1$, $\Lambda_{1,p} = \Lambda_{p,1} = 0.4$ and $\Lambda_{i,i-1} = \Lambda_{i-1,i} = 0.5$, $i = 1, ..., p$.

The choice of the above graphs has been made intentionally to inspect the effect of

sparsity and that of the degree distribution on the nodes of the graph on the performance of the edge selection algorithm. The first five (i.e. Fig. 4.3 (a)-(e)) occur in Yuan and Lin [2007] as well. Note that, the graph corresponding to AR(1), Star and Cluster models have the same number of edges.

However, their degree distributions are vastly different. In AR(1), all nodes have degree two, excepting two nodes at the end, which have degree one. On the other hand the Star model has one node with degree $p-1$ and the rest of the nodes have degree one. For the Cluster models, each cluster has five nodes. This means, the degree of a node is either one or five or six.

The Petersen, Generalized Quadrangle and Cocktail Party graph are strongly regular graphs. Circle is a regular graph where each node has degree 2. The AR(2) and AR(4) models are relatively less sparse with moderate variation in the degree of the nodes.

4.6.2 A comparison of True positives before a fixed proportion of possible False Positives are selected

For each model described, we generate 100 data sets with $p = 10$, $n = 50$ and $p = 30$, $n = 20$. For each data, we record the ceiling of the number of True positives before 5% of the maximum possible False positive edges are added to the active edge set. We report the average of the recorded numbers over the generated 100 data sets.

The results are presented in Table 4.1 From the table, it appears that all methods perform well when the model is sparse and degree distribution is almost uniform. As for example for AR(1) and circle models most of the true edges are selected by the 5% of the FP edges gets in the active set. In contrast, even though the star and the cluster models have same level of sparsity, the performance is markedly bad. This difference in the performance is due to the non-uniform degree distribution of these two models.

Petersen graphs and AR(2) have similar level of sparsity. The degree distribution of Petersen graph is uniform whereas that of AR(2) is nearly uniform. So as expected the algorithms perform well for these graphs. In fact, their performance for the Petersen

Table 4.1 Average number of true positives before 5% of false positives.

(p, n)	Model	$\#\mathcal{E}$	Max FP	5% of max FP	Average number of true positives					
					MB			SPACE		
					ES	AND	OR	NULL	SW	DEW
$(10, 50)$	AR(1)	9	36	2	8.91	8.93	8.89	8.93	8.79	8.89
	AR(2)	17	28	2	8.19	8.36	8.04	8.12	7.63	8.08
	AR(4)	30	15	1	7.47	7.45	7.63	7.18	7.00	7.06
	Star	9	36	2	3.31	3.09	3.25	2.67	2.87	2.93
	Cluster	9	36	2	2.96	3.07	2.96	2.79	2.85	2.86
	Petersen	15	30	2	10.95	11.16	10.8	11.05	10.76	10.71
	Circle	10	35	2	8.94	9.06	8.92	9.05	8.63	8.81
$(10, 200)$	AR(1)	9	36	2	9.00	9.00	9.00	9.00	9.00	9.00
	AR(2)	17	28	2	12.38	13.40	12.17	12.32	12.36	12.29
	AR(4)	30	15	1	13.38	13.65	13.11	13.14	13.27	13.37
	Star	9	26	2	8.11	7.99	8.14	7.89	8.22	8.27
	Cluster	9	26	2	7.85	7.78	7.87	7.74	7.82	7.81
	Petersen	15	30	2	14.97	14.98	14.96	14.97	14.97	14.98
	Circle	10	35	2	10.00	10.00	10.00	10.00	10.00	10.00
$(30, 20)$	AR(1)	29	406	21	28.34	27.92	28.09	28.76	28.27	28.66
	AR(2)	57	378	19	15.34	15.40	15.08	14.70	14.14	14.48
	AR(4)	110	325	17	14.89	14.59	14.53	14.39	14.09	14.11
	Star	29	406	21	6.70	4.42	6.27	3.79	4.56	4.61
	Cluster	29	406	21	5.40	4.90	5.15	4.40	4.72	4.54
	15-CPG	420	15	1	0.70	0.74	0.81	0.94	0.99	1.08
	Circle	30	405	21	29.37	28.88	29.26	29.74	29.75	29.76
$(15, 50)$	GQ(2,2)	45	60	3	36.26	36.07	34.94	36.32	35.28	35.58
$(15, 100)$	GQ(2,2)	45	6	3	44.68	44.38	44.61	44.90	44.62	44.77
$(30, 100)$	15-CPG	420	15	1	0.00	0.00	0.00	0.00	0.00	0.00

graph is slightly better than that of AR(2). In terms of sparsity, GQ(2,2) is roughly equivalent to AR(4). However, GQ(2,2) is a strongly regular graph whereas AR(4) is not. Table 4.1 shows that the performance on GQ(2,2) is better than for AR(4). However, sparsity is important. The graph 15-CPG is strongly regular and almost complete. The simulation study shows that none of the procedures fair well in this case.

When $n < p$, other then the AR(1) and the circle, none of the models can be accurately selected by any of the procedures. Thus in this situation, both sparsity and the degree distribution is important. We observe that MB $-$ OR and ES performs better than the others for Star model. For all other models, all methods seem to be more or less equivalent.

Table 4.2 Models with $p = 10$ nodes, with the methods discussed in section 4.5.

		n = 50				n = 200			
		TP	FP	SE	MC	TP	FP	SE	MC
AR(1)	MB − AND	8.77	1.02	0.11	0.917	9.00	0.52	0.08	0.965
	MB − OR	8.95	4.41	0.27	0.763	9.00	2.77	0.20	0.840
$\#(\mathcal{E}) = 9$	SPACE.BIC	9.00	7.15	0.28	0.668	9.00	5.81	0.23	0.714
	ES.CV$_1$	9.00	3.81	0.21	0.793	9.00	2.24	0.16	0.867
	ES.CV$_{\text{min}}$	9.00	11.52	0.49	0.546	9.00	11.62	0.52	0.544
	ES − OLS	8.96	2.14	0.28	0.869	9.00	0.62	0.17	0.959
	ES − IPF	8.95	1.69	0.21	0.892	9.00	0.48	0.17	0.968
AR(2)	MB − AND	7.49	1.97	0.16	0.441	14.58	2.13	0.13	0.784
	MB − OR	12.13	6.68	0.30	0.467	16.30	5.89	0.26	0.726
$\#(\mathcal{E}) = 17$	SPACE.BIC	14.15	9.00	0.46	0.496	16.94	8.84	0.31	0.667
	ES.CV$_1$	12.92	6.11	0.29	0.532	16.88	6.34	0.24	0.744
	ES.CV$_{\text{min}}$	15.53	14.73	0.41	0.400	16.99	17.86	0.42	0.419
	ES − OLS	13.08	7.68	0.59	0.482	16.64	6.53	0.45	0.723
	ES − IPF	11.13	5.66	0.49	0.454	16.70	6.08	0.40	0.742
AR(4)	MB − AND	3.47	0.44	0.07	0.144	12.42	0.68	0.10	0.383
	MB − OR	10.05	2.28	0.19	0.207	19.90	3.11	0.19	0.430
$\#(\mathcal{E}) = 30$	SPACE.BIC	7.59	1.15	0.17	0.210	21.81	4.97	0.34	0.380
	ES.CV$_1$	10.12	2.03	0.22	0.215	22.41	4.56	0.26	0.426
	ES.CV$_{\text{min}}$	20.34	7.55	0.35	0.170	28.98	11.53	0.25	0.310
	ES − OLS	13.83	3.98	0.45	0.189	28.34	10.48	0.36	0.337
	ES − IPF	6.13	1.00	0.25	0.178	22.92	7.04	0.52	0.294
Star	MB − AND	0.74	0.23	0.06	0.209	2.31	0.05	0.02	0.458
	MB − OR	3.51	1.61	0.24	0.435	8.04	0.85	0.16	0.874
$\#(\mathcal{E}) = 9$	SPACE.BIC	0.76	0.30	0.07	0.201	8.29	3.40	0.22	0.754
	ES.CV$_1$	1.00	0.47	0.13	0.221	7.54	1.29	0.17	0.808
	ES.CV$_{\text{min}}$	4.75	6.48	0.51	0.322	8.92	12.67	0.49	0.512
	ES − OLS	3.10	2.31	0.41	0.345	8.12	2.47	0.36	0.786
	ES − IPF	1.19	0.83	0.21	0.211	8.17	2.54	0.35	0.786
Cluster	MB − AND	0.56	0.15	0.05	0.186	2.06	0.13	0.04	0.419
	MB − OR	2.51	1.94	0.24	0.301	6.42	1.31	0.17	0.718
$\#(\mathcal{E}) = 9$	SPACE.BIC	0.97	0.50	0.13	0.211	6.70	2.14	0.23	0.690
	ES.CV$_1$	0.56	0.16	0.07	0.184	6.33	0.91	0.13	0.738
	ES.CV$_{\text{min}}$	4.60	6.96	0.59	0.291	8.83	12.33	0.48	0.512
	ES − OLS	2.67	2.03	0.32	0.314	7.43	2.45	0.33	0.732
	ES − IPF	0.67	0.37	0.14	0.171	7.71	3.18	0.38	0.718
Petersen	MB − AND	4.95	0.83	0.10	0.426	14.18	0.38	0.06	0.940
	MB − OR	9.89	3.60	0.25	0.555	14.85	2.17	0.20	0.892
$\#(\mathcal{E}) = 15$	SPACE.BIC	12.03	4.53	0.43	0.636	15.00	4.47	0.29	0.810
	ES.CV$_1$	11.35	3.05	0.23	0.662	14.99	2.16	0.20	0.900
	ES.CV$_{\text{min}}$	14.04	12.48	0.40	0.498	15.00	14.45	0.40	0.514
	ES − OLS	12.60	5.19	0.56	0.644	14.96	0.93	0.15	0.953
	ES − IPF	9.51	2.71	0.35	0.576	14.96	1.31	0.18	0.936
Circle	MB − AND	9.99	2.62	0.17	0.856	10.00	1.58	0.12	0.908
	MB − OR	10.00	7.30	0.32	0.676	10.00	5.05	0.25	0.754
$\#(\mathcal{E}) = 10$	SPACE.BIC	10.00	8.98	0.25	0.626	10.00	8.21	0.19	0.648
	ES.CV$_1$	10.00	6.54	0.21	0.701	10.00	5.29	0.21	0.745
	ES.CV$_{\text{min}}$	10.00	11.60	0.47	0.556	10.00	10.96	0.42	0.572
	ES − OLS	9.94	2.34	0.28	0.865	10.00	0.64	0.24	0.961
	ES − IPF	9.96	2.38	0.32	0.865	10.00	0.77	0.20	0.953

Table 4.3 Models with $p = 15$ nodes, with the methods discussed in section 4.5.

		n = 50				n = 200			
		TP	FP	SE	MC	TP	FP	SE	MC
AR(1)	MB − AND	13.84	1.67	0.12	0.930	14.00	0.63	0.07	0.975
	MB − OR	13.94	9.06	0.54	0.737	14.00	4.89	0.27	0.837
$\#(\mathcal{E}) = 14$	SPACE.BIC	14.00	13.76	0.45	0.654	14.00	11.51	0.32	0.692
	ES.CV_1	14.00	7.39	0.27	0.776	14.00	5.55	0.28	0.820
	ES.$CV_{\min}$	14.00	20.09	0.68	0.566	14.00	22.17	0.77	0.541
	ES − OLS	13.89	2.32	0.19	0.909	14.00	0.43	0.10	0.983
	ES − IPF	13.91	2.62	0.29	0.900	14.00	0.58	0.12	0.977
AR(2)	MB − AND	11.36	3.91	0.29	0.460	23.11	4.56	0.21	0.791
	MB − OR	18.51	13.88	0.62	0.480	25.88	13.57	0.47	0.708
$\#(\mathcal{E}) = 27$	SPACE.BIC	20.69	16.14	0.73	0.512	26.92	16.42	0.51	0.698
	ES.CV_1	20.64	15.66	0.60	0.518	26.72	15.88	0.48	0.700
	ES.$CV_{\min}$	24.31	35.07	0.88	0.397	26.97	41.40	0.80	0.429
	ES − OLS	19.74	15.13	1.00	0.499	26.45	9.43	0.37	0.791
	ES − IPF	14.72	7.26	0.72	0.486	26.45	10.66	0.49	0.771
AR(4)	MB − AND	4.82	1.03	0.13	0.169	18.48	2.04	0.19	0.419
	MB − OR	15.16	6.98	0.45	0.216	30.44	10.07	0.47	0.437
$\#(\mathcal{E}) = 50$	SPACE.BIC	8.59	2.13	0.25	0.220	26.20	6.34	0.54	0.441
	ES.CV_1	12.90	3.81	0.40	0.258	37.52	15.59	0.51	0.466
	ES.$CV_{\min}$	28.82	19.17	0.87	0.228	47.02	37.70	0.56	0.323
	ES − OLS	12.29	4.57	0.54	0.221	41.34	27.11	1.47	0.350
	ES − IPF	6.81	0.92	0.13	0.228	23.93	7.52	1.30	0.373
Star	MB − AND	1.70	0.48	0.09	0.277	5.03	0.09	0.03	0.566
	MB − OR	7.76	4.42	0.48	0.537	13.55	1.68	0.26	0.916
$\#(\mathcal{E}) = 14$	SPACE.BIC	1.40	0.95	0.25	0.206	13.48	6.95	0.38	0.761
	ES.CV_1	3.93	1.46	0.27	0.408	13.36	3.25	0.31	0.856
	ES.$CV_{\min}$	9.40	16.02	0.91	0.393	13.94	24.17	0.78	0.516
	ES − OLS	5.33	3.09	0.50	0.434	13.05	2.97	0.34	0.850
	ES − IPF	3.02	2.17	0.39	0.301	13.41	3.74	0.37	0.843
Cluster	MB − AND	0.63	0.31	0.06	0.150	3.35	0.13	0.04	0.452
	MB − OR	3.19	3.97	0.36	0.248	10.01	2.79	0.31	0.711
$\#(\mathcal{E}) = 14$	SPACE.BIC	0.67	0.43	0.11	0.144	10.33	3.23	0.28	0.712
	ES.CV_1	0.76	0.42	0.15	0.160	10.65	2.48	0.27	0.754
	ES.$CV_{\min}$	6.13	11.72	0.97	0.280	13.63	26.03	0.77	0.482
	ES − OLS	2.44	1.48	0.26	0.283	11.55	3.39	0.33	0.767
	ES − IPF	0.10	0.00	0.00	0.025	11.62	5.13	0.47	0.718
GQ(2,2)	MB − AND	38.30	4.07	0.22	0.790	45.00	1.89	0.14	0.964
	MB − OR	43.51	12.53	0.46	0.752	45.00	8.12	0.30	0.856
$\#(\mathcal{E}) = 45$	SPACE.BIC	44.34	23.39	0.86	0.616	45.00	15.62	0.52	0.741
	ES.CV_1	44.36	12.78	0.37	0.768	45.00	10.00	0.43	0.826
	ES.$CV_{\min}$	44.56	25.96	0.56	0.588	45.00	27.95	0.73	0.574
	ES − OLS	43.19	7.35	0.43	0.829	44.98	0.69	0.14	0.986
	ES − IPF	43.61	8.81	0.37	0.814	45.00	0.68	0.15	0.987
Circle	MB − AND	14.97	1.73	0.15	0.937	15.00	0.73	0.09	0.973
	MB − OR	15.00	8.54	0.547	0.759	15.00	4.92	0.29	0.844
$\#(\mathcal{E}) = 15$	SPACE.BIC	15.00	12.30	0.41	0.689	15.00	9.76	0.31	0.735
	ES.CV_1	15.00	7.85	0.34	0.774	15.00	5.73	0.28	0.823
	ES.$CV_{\min}$	15.00	20.78	0.76	0.568	15.00	20.90	0.613	0.566
	ES − OLS	14.95	1.65	0.16	0.938	15.00	0.49	0.13	0.981
	ES − IPF	14.98	2.05	0.253	0.926	15.00	0.43	0.12	0.984

Table 4.4 $n = 20$, $p = 30$.

Method	Model	TP	FP	SE	MC	Model	TP	FP	SE	MC
MB − AND	AR(1)	26.52	22.47	0.89	0.678	Cluster	2.66	17.48	0.88	0.058
MB − OR		28.30	131.91	2.70	0.337		11.42	115.68	2.88	0.060
SPACE.BIC	$\#(\mathcal{E})$	28.72	47.86	1.13	0.571	$\#(\mathcal{E})$	0.30	0.75	0.28	0.043
ES.CV$_1$	29	28.57	27.40	0.54	0.684	29	0.16	0.15	0.11	0.048
ES.CV$_{\min}$		28.68	47.71	0.68	0.571		1.88	6.46	1.15	0.089
ES − OLS		27.47	11.28	0.51	0.805		0.39	0.45	0.08	0.070
MB − AND	AR(2)	10.55	21.89	1.00	0.163	15-CPG	20.64	2.90	0.15	-0.116
MB − OR		29.83	122.7	2.89	0.141		128.12	8.06	0.23	-0.091
SPACE.BIC	$\#(\mathcal{E})$	2.61	2.31	0.60	0.127	$\#(\mathcal{E})$	0.79	0.59	0.16	-0.122
ES.CV$_1$	57	5.89	6.32	0.92	0.177	420	1.31	0.73	0.21	-0.122
ES.CV$_{\min}$		20.63	38.76	2.23	0.255		21.65	5.47	0.39	-0.236
ES − OLS		4.60	4.28	0.84	0.166		1.13	1.19	0.18	-0.192
MB − AND	AR(4)	8.11	14.30	0.77	0.058	Circle	27.62	26.21	0.94	0.659
MB − OR		38.15	93.19	2.57	0.057		29.35	135.49	3.01	0.336
SPACE.BIC	$\#(\mathcal{E})$	0.89	0.56	0.19	0.048	$\#(\mathcal{E})$	29.71	45.70	0.95	0.587
ES.CV$_1$	110	1.21	1.01	0.34	0.048	30	29.52	31.88	0.56	0.659
ES.CV$_{\min}$		11.59	15.57	1.49	0.103		29.63	48.54	0.72	0.573
ES − OLS		1.10	0.43	0.11	0.064		28.96	11.34	0.54	0.819
MB − AND	Star	2.57	14.72	0.86	0.067					
MB − OR		11.87	107.25	3.05	0.081					
SPACE.BIC	$\#(\mathcal{E})$	0.16	0.28	0.11	0.038					
ES.CV$_1$	29	0.54	0.63	0.31	0.082					
ES.CV$_{\min}$		4.65	12.74	1.37	0.164					
ES − OLS		1.69	0.63	0.16	0.194					

4.6.3 Edge Selection with proposed Cross Validation methods

We cross validate edge selection according to the methods described in Section 4.5.2. More specifically for ES.CV$_e$, we take $e = 1$ (denoted ES.CV$_1$). The iterative proportional fitting algorithm requires $n > p$. Thus ES.IPF cannot be used when $p > n$.

We compare our method with MB − OR, MB − AND and the SPACE method after cross-validation. We use Bayesian information criterion (BIC) to cross-validate SPACE [Peng et al., 2009]. The method is denoted by SPACE.BIC.

Both MB − OR and MB − AND are cross-validated for each neighborhood separately using a method similar to ES.CV$_1$. Note that, our cross-validation method for MB − OR and MB − AND defers from the methods proposed by Meinshausen and Bülmann [2006] and Yuan and Lin [2007]. However, our simulation studies show that our method is comparable to their performance for finite sample sizes.

The average of the number of true positive (TP) and false positive (FP) edges selected over 100 simulations are presented in Tables 4.2, 4.3 and 4.4. We also provide the standard deviation of the number of false positives (SE) and the average Matthews correlation coefficient (MC).

Using Matthews correlation as a measurement for performance, we first look at the case when $n > p$.

For AR(1) and Circle, where the degree of the nodes and the number of edges are low, all methods seem to be able to select most of the correct edges. This would imply that the difference in Matthews correlation of each method would depend on the number of false positives. We also observe that in such cases, MB − AND, ES − OLS and ES − IPF have very low false positives.

Keeping the number of edges low, and setting of the degree of some of the nodes high, like Star and Cluster model, all methods do not perform well when $n = 50$. However, the performances of all methods except MB − AND improve dramatically when $n = 200$. Comparing between methods, Table 4.2 and 4.3 shows that for these two models, MB − OR is better than all the other methods excepting ES − OLS for Cluster model when $p = 15$.

For AR(4) and GQ(2,2) graph, where the number of edges is higher, it seems that the structure of the graph plays a significant part in performance. In particular, all methods perform significantly better in strongly regular graph such as GQ(2,2). A similar conclusion can be drawn when the number of edges is lower, such as AR(2) and Petersen graph. For all of these four models when $n = 50$, ES.CV$_1$ has relatively better performance.

For the case when $n < p$. the degree of nodes and the number of edges are important as well. Similar to the case when $n > p$, all methods are able to select most of the correct edges for AR(1) and Circle model and have very bad performances for Cluster and Star model. We also note that MB − AND no longer performs well in AR(1) and Circle and MB − OR no longer perform the best for Star and Cluster model. Instead, ES − OLS now outperforms all other methods for all these four models, AR(1), Circle, Star and

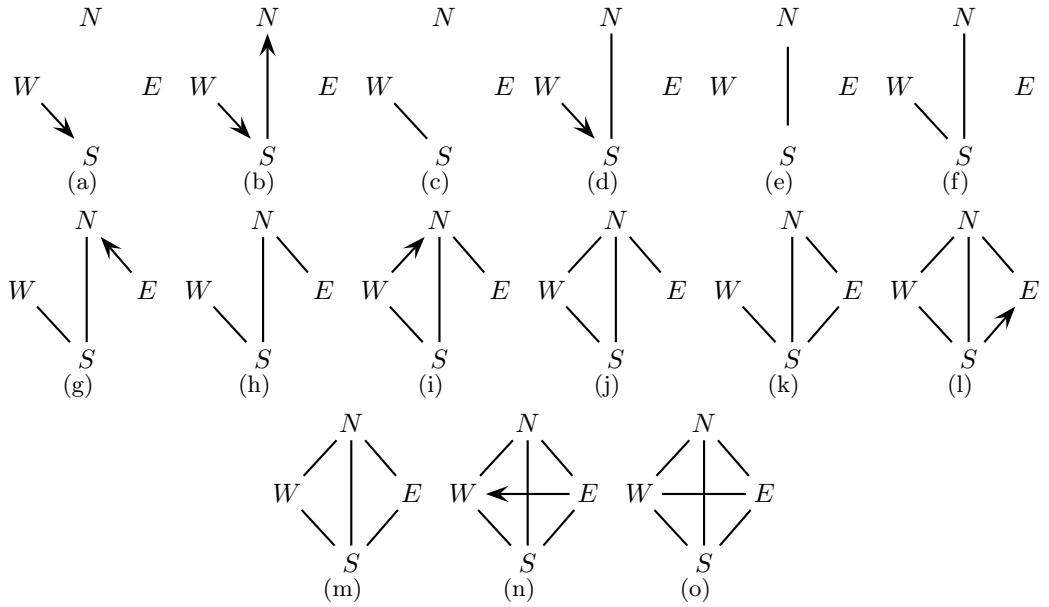


Figure 4.4 A comparison of various model selection methods on the Cork-borings data. MB in succession selects $(a, b, d, f, g, h, i, j, l, m, n, o)$. For MB methods, the path of MB-AND is (e, f, h, j, m, o) and the path of MB-OR is (c, f, h, j, m, o), The paths of ES and SPACE are both (c, f, h, k, m, o). Upon cross validation, ES.CV$_1$, SPACE.BIC and MB $-$ OR pick (m), while MB $-$ AND pick (j).

Cluster. We also observe that ES.CV$_1$ seems to select too little edges. In fact, ES.CV$_{\min}$ is now performing better than ES.CV$_1$ for all models except AR(1), Circle and 15-CPG. We conclude that for Edge selection, we should either use ES.CV$_{\min}$ or ES $-$ OLS for $n < p$. Note that taking the minimum does not help neighborhood selection as both MB $-$ AND and MB $-$ OR are selecting too many edges.

4.7 Application to real data sets

4.7.1 Cork borings data

We consider the cork borings data [Mardia et al., 1979, page 11] first. The data set consists of the weights of cork deposits in four cardinal directions (ie. North, South, East and West) obtained from 28 cork trees. In figure 4.4 we present the full paths of MB,

MB − AND, MB − OR,SPACE and the proposed ES method.

In the path for MB we display the edges selected in any of the four neighborhoods as the tuning parameter of the LASSO (denoted τ) decreases. A directed edge, as for example, $W \to S$ implies that at this point in the path, W is in the neighborhood of S, but the S is still not a neighbor of W. When W and S are neighbors of each other we draw an undirected edge between them.

The path of MB turns out to be $(a, b, d, f, g, h, i, j, l, m, n, o)$. The MB − OR and MB − AND would respectively in succession choose (e, f, h, j, m, o) and (c, f, h, j, m, o). These two paths are only different at their starting points, but with cross-validation MB − OR would pick model (m), where as MB − OR chooses model (j). The path for ES and SPACE are exactly the same, both are (c, f, h, k, m, o). ES.CV$_1$ and SPACE.BIC both pick model (m).

4.7.2 Mathematics examination marks data

Next we consider the mathematics examination marks data from Mardia et al. [1979]. This data set consists of grades of 88 students in 5 subjects namely Mechanics, vectors, algebra, analysis and statistics. The paths for MB − AND, MB − OR, SPACE and ES has been displayed in Figure 4.5. Due to space constraint we do not present the full path for MB here. It turns out that the paths for MB − OR $((a, e, h, l, m, o, p, r, u, v))$, MB − AND $((b, f, h, j, n, o, p, r, u, v))$, SPACE $((c, d, g, j, m, o, p, q, t, v))$ and ES $((b, e, i, k, n, o, p, s, u, v))$ have little in common. After cross-validation SPACE.BIC picks (p), while MB − AND, MB − OR and ES.CV$_1$ pick (o).

4.7.3 Application to isoprenoid pathways in Arabidopsis thaliana

We apply our method on a data involving the gene-expression patterns for isoprenoid pathways in Arabidopsis thaliana. The pathways were monitored using 118 micro-arrays, with a regulatory network of 39 genes. These 39 genes can be split into two distinct pathway, namely, the MVA pathway with 13 genes and the MEP pathway with 19 genes. The rest of the genes were from the Mitochondrion.

For MEP pathway, it is responsible for synthesis of isoprenes, carotenoids and the side chains of chlorophyll and platoquinone. MVA pathway, on the other hand, is responsible for the synthesis of sterols, sesquiterpenes and the side chain of ubiquinone.[Wille et al., 2004]. Interaction between these two pathways have been reported by several authors (see Laule et al. [2003],Rodriguez-Concepcion et al. [2004], among others). The scientific question is to find out the genes through which these pathways interact.

Wille et al. [2004] applied two different graphical modeling approaches to discover the mode of interaction between isoprenoid pathways in Arabidopsis thaliana. The first approach uses conventional Gaussian Graphical model method (GGM) with backward selection with BIC. This was carried out using MIM 3.1 [MIM, 2009]. Their second approach was a modified GGM method based on frequentist hypothesis testing and thresholding. This modified GGM selected 31 edges, many of which agreed with concurrent experimental findings, prior knowledge and estimated absolute pairwise correlations.

We applied ES, MB − OR and MB − AND on this data set. No information about the directed nature of the MVA and MEP pathways were included in the procedure. We endeavored to choose an undirected graph. For ES, $e = 1$ selects 78 edges, which are too many to interpret. We choose $e = 2$ instead, which selects a model with 32 edges, comparable in count with 31 edges Wille et al. [2004] chose. We denote our method as ES.CV_2.

The resulting model can be found in Figure 4.6. Our selected model shows that DXPS2, DXR, MCT, MECPS and HDS are nearly fully connected, which is similar to the findings of the modified GGM approach. We also found more edges between MEP pathway and gene HMGR1. This complies with the expectation that HMGR1 is an important gene in the communication network between MEP and MVA pathways (Wille et al. [2004]).

The connections between AACT2, MK, MPDC1 and FPPS2 were found to be weak. However, we managed to detect quite a few edges between AACT2 and FPPS2 which, despite being highly correlated, are not selected by the modified GGM approach. Another observation is that ES.CV_2 did not detect almost any edges on two directed pathways,

which in view of the fact that these edges are supposed to be actually directed, to a great extent validates our method.

We found 18 common edges among the first 32 edges for MB-AND and MB-OR. Compared to ES, both MB-AND and MB-OR seem to show more connections within the MEP pathway. It is also observed that $MB - OR$ show stronger connection between MEP pathway and HMGR1 than $MB - AND$.

4.8 Discussion

In this chapter, we propose a symmetric way of selecting edges in an Undirected Graph. Our method is based on the geometric application of the least angle regression and group LARS proposed by Efron et al. [2004] and Yuan and Lin [2006] respectively. We show how this approach of Edge Selection can be interpreted as selecting variable based on selecting varaibles based on the maximality of $\mathcal{C}^2$. Furthermore, with step-wise linearity of the paramter estimates of the Edge Selection method, this allows an efficient method to select edges from the model.

We also identify a parameter, $\mathcal{M}_0$, which we use to compare between models on the selection path. This parameter is piecewise linear with respect to our parameter estimates and has certain nice properties which are similar to the regularization parameter used in LASSO. We also show how we can select $\mathcal{M}_0$ based on Cross Validation.

Theoetical questions such as consistency are still open. However, simulation studies involving a range of models shows that Edge Selection has comparable results with other known methods.

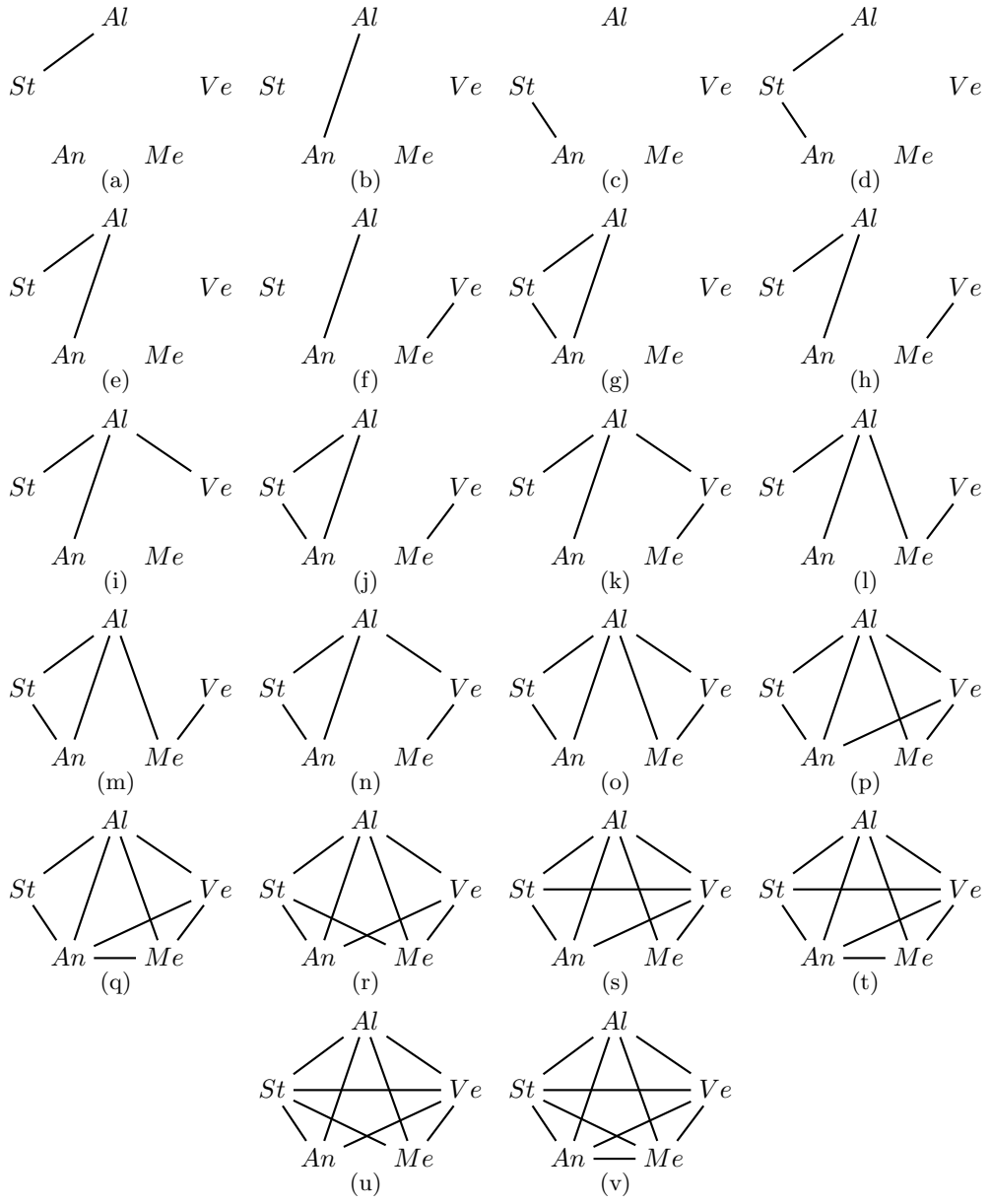


Figure 4.5 Results for the Mathematics marks dataset. The paths of MB – OR is $(a, e, h, l, m, o, p, r, u, v)$, for MB – AND is $(b, f, h, j, n, o, p, r, u, v)$, for SPACE is $(b, e, i, k, n, o, p, s, u, v)$ and for ES is $(c, d, g, j, m, o, p, q, t, v)$. Cross-validated MB – OR, MB – AND and ES.CV$_1$ all pick model (o), while SPACE.BIC chooses model (p).

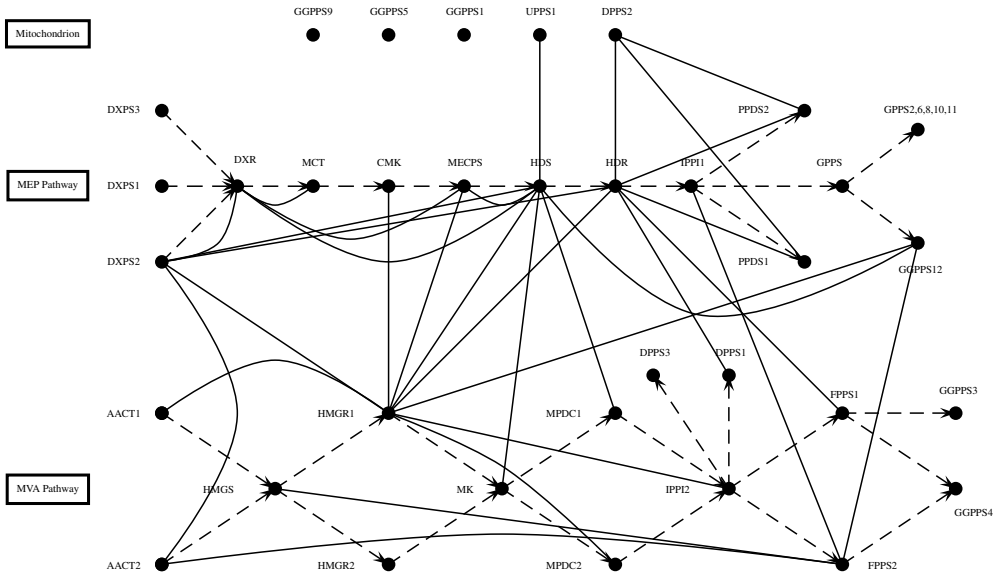


Figure 4.6 The directed arrows represent the underlying pathway in Arabidopsis thaliana. The undirected Edges are selected by $ES.CV_2$

CHAPTER 5

LASSO with known Partial Information

5.1 Introduction

Two of the most important goals in building a regression model are to achieve good model estimation and model selection. A good model estimation is achieved when there is high prediction accuracy. One measure of prediction accuracy that is commonly used is the mean squared errors, which in turn depends on the bias and variance of the estimated model. A good model selection method is able to select a model that is as close to the true model as possible. Model selection is especially important because interpretability of the model becomes difficult when we are given a dataset with many predictors and the underlying model has a sparse representation.

There are many methods to estimate a linear regression model. Traditional methods such as ordinary least squares and ridge regression can be used to estimate $\hat{\boldsymbol{\beta}}$. However, as they always keep all the predictors in the model, they cannot produce a parsimonious

model. In order to select a parsimonious model, methods such as stepwise regression or LASSO have to be used. In particular, LASSO is an extremely popular tool for performing both model estimation and selection.

LASSO is popular because it is known to be computational efficient, and can achieve good model selection and estimation. Its biggest advantage is that it does not require the user to search the space of all models, which can be extremely large. LASSO is widely used in many applications including genetics, networks and signals. LASSO can handle sparse models when the number of observations is less than the number of covariates. This has been extremely useful in many examples.

In most of the applications encountered in real life, some information about the underlying model is known. As for example, in genetic studies, some of the relevant genes may be available from the experiment or background knowledge. In most cases, the question of interest is to find the additional covariates in the model for the response. From an intuitive point of view, it is clear that such available information should be included in the model. Standard LASSO ignore this information and tries to select a model from scratch. This clearly is not desired or particularly efficient. If one insists on using standard LASSO, one way to ensure that all variables known to be in the true model are selected is to choose an appropriate shrinkage. The shrinkage can be chosen in such a way so that all such variables are included in the model. In most cases, the chosen model would be extremely large and may contain variables which are not particularly relevant for the response. This is because even though some variables are in the true model, their observed correlation with the response may not be particularly high and LASSO would choose these variables much later on the solution path.

One natural modification of standard LASSO in this case is the so called "first regression, then LASSO" procedure. In this method, one first regresses $\mathbf{Y}$ on the variables known from background knowledge and then tries to explain the residual from this regression with additional covariates using an ordinary LASSO procedure. This procedure is not optimal because the estimate of the parameter vector obtained during the first regression step remain invariant in the second LASSO step.

In this chapter, we propose a solution of this problem. We propose a special case of the weighted LASSO to force these known variables to always be on the solution path. In that way, these variables are always selected in the model. We call this method the Partial LASSO, or PLASSO. Our method is based on LASSO but it only put L_1 constraint on those available variables which are not known to be in the true model from background information. We minimize a quadratic loss in explaining the response with all the available variables under the above constraint.

The PLASSO is a convex problem so Partial Least Angle Regression (PLARS) algorithm can be devised to quickly compute the whole model selection path. We show that this PLARS has many desirable properties. Similar to the "first regression then LASSO" procedure, it keeps the known variables in the model and then tries to select additional variables to explain part of the variation in the response not explained by those known variables. However, unlike the former, it estimates the whole parameter vector optimally at every step. Additionally, we show that the proposed method is estimation consistent under usual assumptions.

We also investigate the selection consistency of our proposed PLASSO procedure. We find conditions under which PLASSO is selection consistent. These conditions are compared with those for standard LASSO. Using simulation studies, we show that in many cases, when LASSO is not selection consistent, PLASSO may be selection consistent. However, surprisingly, in many cases, it turns out that even though LASSO is selection consistent, PLASSO is not. This is a strange observation which implies in many cases throwing away background about the model might actually be beneficial. In any case, our results show that selection consistency of standard LASSO and PLASSO does not imply each other.

5.2 Notations and Assumptions

We start by introducing the notations used in this section and describe our assumptions. Suppose $\mathbf{Y} = (Y_1, ..., Y_n)^T$ is $n \times 1$ response vector and

$$\mathbf{X} = \begin{pmatrix} X_{11} & X_{12} & \dots & X_{1p} \\ X_{21} & X_{22} & \dots & X_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ X_{n1} & X_{n2} & \dots & X_{np} \end{pmatrix}$$

is the matrix of covariates.

Further, suppose $\mathbf{x}_i = (X_{i1}, ..., X_{ip})$., and $\mathbf{X}_j = (X_{1j}, ..., X_{nj})^T$ respectively denote the *ith* row and *jth* column of, that is $(\mathbf{x_1^T}, ..., \mathbf{x_n^T})^\mathbf{T} = (\mathbf{X_1}, ..., \mathbf{X_p})$. Without loss of generality, we assume that $\mathbf{Y}$ is centered and each column vector $\mathbf{X}_j$ is standardized. That is,

$$\sum_{i=1}^{n} Y_i = 0. \ \sum_{i=1}^{n} X_{ij}^2 = 1, \ \sum_{i=1}^{n} X_{ij} = 0.$$

We consider the model

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{5.2.1}$$

where $\boldsymbol{\epsilon}$ is the vector of errors which is normally distributed with mean $\mathbf{0}$ and variance $\sigma^2\mathbf{I}_p$. Suppose that there is an unknown set $\mathcal{A}_3 \subseteq \{1, ...p\}$ such that $\boldsymbol{\beta}_{\mathcal{A}_3} = \mathbf{0}$. That is, not all columns of $\mathbf{X}$ is in the true model. Further, suppose that from the background information, it is known that some columns of $\mathbf{X}$ are in the true model. Let us collect these indices of these columns in $\mathcal{A}_1 \subseteq \{1, ...p\}$. What we mean is that $\boldsymbol{\beta}_{\mathcal{A}_1}$ is known to be nonzero. However, the actual value of $\boldsymbol{\beta}_{\mathcal{A}_1}$ is not known and has to be estimated from the data. Let $\mathcal{A}_2 = \{1, ...p\} \backslash \mathcal{A}_1 \cup \mathcal{A}_3$. Thus, the elements in $\boldsymbol{\beta}_{\mathcal{A}_2}$ are nonzero but we do not know that. By our definition, $\boldsymbol{\beta} = (\boldsymbol{\beta}_{\mathcal{A}_1}, \boldsymbol{\beta}_{\mathcal{A}_2}, \boldsymbol{\beta}_{\mathcal{A}_3})$ and let the cardinality of $\mathcal{A}_1$, $\mathcal{A}_2$ and $\mathcal{A}_3$ be a_1, a_2 and a_3 respectively. We further denote $\boldsymbol{\beta}_{\mathcal{A}} = (\boldsymbol{\beta}_{\mathcal{A}_1}, \boldsymbol{\beta}_{\mathcal{A}_2})$ and

$\boldsymbol{\beta}_{\mathcal{A}'} = (\boldsymbol{\beta}_{\mathcal{A}_2}, \boldsymbol{\beta}_{\mathcal{A}_3})$. Note that $\boldsymbol{\beta}_{\mathcal{A}}$ groups all the regression coefficients in the true model whereas $\mathcal{A}' = \{\mathcal{A}_2, \mathcal{A}_3\}$ is the set of all coefficients for which there is no background knowledge. Our goal is to find out the elements of $\boldsymbol{\beta}_{\mathcal{A}'}$ which are nonzeros.

Following the setup of Knight and Fu [2000], Zhao and Yu [2006], we assume the following regularity conditions :

(A1) For a positive definite matrix $\mathbf{C}$,

$$\frac{1}{n}\mathbf{X}^T\mathbf{X} = \mathbf{C} \to \Sigma, \; as \; n \to \infty. \tag{5.2.2}$$

(A2) We assume that

$$\frac{1}{n} max_{1 \le i \le n} \mathbf{x_i}\mathbf{x_i}^T \to 0, \; as \; n \to \infty. \tag{5.2.3}$$

(A3) Further, we assume that

$$\mathbf{W} = \frac{\mathbf{X}^T\boldsymbol{\epsilon}}{\sqrt{n}} \to_d \mathcal{W}, \; as \; n \to \infty \tag{5.2.4}$$

where $\mathcal{W}$ has a $N(\mathbf{0}, \sigma^2\boldsymbol{\Sigma})$ distribution.

We note that $\mathbf{C}$ is also the sample covariance matrix. Moreover, because the columns of $\mathbf{X}$ are standardized, each diagonal element of $\mathbf{C}$ is equal to one.

Corresponding to $\boldsymbol{\beta} = (\boldsymbol{\beta}_{\mathcal{A}_1}, \boldsymbol{\beta}_{\mathcal{A}_2}, \boldsymbol{\beta}_{\mathcal{A}_3})$, the data matrix and response vector can be written as $\mathbf{X} = (\mathbf{X}_{\mathcal{A}_1}, \mathbf{X}_{\mathcal{A}_2}, \mathbf{X}_{\mathcal{A}_3})$ and $\mathbf{Y} = (\mathbf{Y}_{\mathcal{A}_1}, \mathbf{Y}_{\mathcal{A}_2}, \mathbf{Y}_{\mathcal{A}_3})$. Finally, the sample covariance matrices $\mathbf{C}$ can be divided accordingly into their sub-matrices as well. We specifically denote

$$\mathbf{C} = \begin{pmatrix} \mathbf{C}_{11} & \mathbf{C}_{12} & \mathbf{C}_{13} \\ \mathbf{C}_{21} & \mathbf{C}_{22} & \mathbf{C}_{23} \\ \mathbf{C}_{31} & \mathbf{C}_{32} & \mathbf{C}_{33} \end{pmatrix}, \mathbf{C}_{\mathcal{A}} = \mathbf{C}_{\mathcal{A}\mathcal{A}} = \begin{pmatrix} \mathbf{C}_{11} & \mathbf{C}_{12} \\ \mathbf{C}_{21} & \mathbf{C}_{22} \end{pmatrix},$$

$$\mathbf{C}_{\mathcal{A}'\mathcal{A}'} = \begin{pmatrix} \mathbf{C}_{22} & \mathbf{C}_{23} \\ \mathbf{C}_{32} & \mathbf{C}_{33} \end{pmatrix}, \mathbf{C}_{\mathcal{A}'\mathcal{A}_1} = \begin{pmatrix} \mathbf{C}_{21} \\ \mathbf{C}_{31} \end{pmatrix}, \mathbf{C}_{\mathcal{A}_3\mathcal{A}} = \begin{pmatrix} \mathbf{C}_{31} & \mathbf{C}_{32} \end{pmatrix}.$$

where $\mathbf{C}_{\mathcal{A}_t\mathcal{A}_j} = \mathbf{C}_{tj} = \frac{1}{n}\mathbf{X}_{\mathcal{A}_t}^T\mathbf{X}_{\mathcal{A}_j}$ for $t, j = 1, 2, 3$.

Furthermore, define $\mathbf{C}_{tt|j}$ and $\mathbf{C}_{st|j}$ as

$$\mathbf{C}_{tt|j} = \mathbf{C}_{tt} - \mathbf{C}_{tj}(\mathbf{C}_{jj})^{-1}\mathbf{C}_{jt},\ \mathbf{C}_{st|j} = \mathbf{C}_{st} - \mathbf{C}_{sj}(\mathbf{C}_{jj})^{-1}\mathbf{C}_{jt}.$$

The rest of the chapter is organized as follows. In section 5.3, we introduce our PLASSO method. PLARS algorithm, which is an adaptation of LARS for solving PLASSO problem is described next (Section 5.4). Section 5.5 and 5.6 is dedicated to some asymptotic properties of the proposed method. In particular, we discuss the estimation consistency of PLASSO in section 5.5. The selection consistency and sign consistency of PLASSO is discussed in section 5.6. Finally, in section 5.7, we perform a simulation study on a few examples to evaluate the performance of LASSO and PLASSO.

5.3 PLASSO : LASSO with Known Partial Information

Recall that our prior information does not specify whether the variables corresponding to $\boldsymbol{\beta}_{\mathcal{A}'}$ are in the true model or not. We study a natural solution where the LASSO penalty is imposed only on $\boldsymbol{\beta}_{\mathcal{A}'}$. The Partial LASSO or PLASSO problem estimates $\boldsymbol{\beta}$ by taking

$$\hat{\boldsymbol{\beta}}(t) = arg\min_{\beta}\left\{(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})\right\} \text{ subject to } \quad ||\boldsymbol{\beta}_{\mathcal{A}'}||_1 \leq t. \tag{5.3.1}$$

Note that only $\boldsymbol{\beta}_{\mathcal{A}'}$ is shrunk. Therefore, the variables which are already known to be in the true model will always be selected. The amount of shrinkage t is usually determined by cross validation. Even though, no direct shrinkage of $\boldsymbol{\beta}_{\mathcal{A}_1}$ occurs, it is trivial to note that $\hat{\boldsymbol{\beta}}_{\mathcal{A}_1}$ will not be equal to the unconstrained OLS estimate of $\beta_{\mathcal{A}_1}$. Note that $\hat{\boldsymbol{\beta}}$ also depends on the sample size and the data. For notational convenience, in section 5.3 and 5.4, we suppress its explicit mention. By taking a Lagrange multiplier λ corresponding to the constraint, the PLASSO problem can be reformulated as

$$\hat{\boldsymbol{\beta}}(\lambda) = argmin_{\beta}\left\{(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})\right\} + \lambda||\boldsymbol{\beta}_{\mathcal{A}'}||_1. \tag{5.3.2}$$

The PLASSO problem as described above can be viewed as a weighted LASSO problem (See Chapter 2) where $w_j = 0$ for $j \in \mathcal{A}_1$ and $w_j = 1$ for $j \in \mathcal{A}'$.

PLASSO have several advantages. By construction, all variables that are known to be in the true model are always selected and PLASSO can choose small number of additional variables and thus select a parsimonious model. Further, these known variables will be in the selected model whether or not PLASSO is consistent. From a computational point of view, PLASSO is easy to solve. First of all, from (5.3.1), it is clear that the optimization problem is convex. Thus, most of the algorithms used to solve standard LASSO problem can be adapted to solve PLASSO.

The KKT solution of (5.3.2) can be stated explicitly.

Lemma 5.1 *$\hat{\boldsymbol{\beta}}$ is a solution of* (5.3.2) *if and only if*

$$-\mathbf{X}_j^T(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}) = 0 \; for \; \hat{\beta}_j \in \mathcal{A}_1, \tag{5.3.3}$$

$$-\mathbf{X}_j^T(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}) = -\lambda sign(\hat{\beta}_j) \; for \; \hat{\beta}_j \neq 0, \hat{\beta}_j \notin \mathcal{A}_1, \tag{5.3.4}$$

$$-\mathbf{X}_j^T(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}) \leq \lambda \; for \; \hat{\beta}_j = 0, \hat{\beta}_j \notin \mathcal{A}_1. \tag{5.3.5}$$

Proof: The above equations and inequalities follow directly from the KKT solutions of the (5.3.2). □

From Lemma 5.1, notice that $\hat{\boldsymbol{\beta}}$ satisfies (5.3.3). This would not be true for the standard LASSO solution.

For $t = 0$, (5.3.1) implies that $\hat{\boldsymbol{\beta}}_{\mathcal{A}'}(0) = \mathbf{0}_{\mathcal{A}'}$, where $\mathbf{0}_{\mathcal{A}'}$ is a vector of zeros of length $|\mathcal{A}'| = a_2 + a_3$. Thus, for $t = 0$, the estimate of $\hat{\boldsymbol{\beta}}_{\mathcal{A}_1}$ is obtained as

$$\begin{aligned} \hat{\boldsymbol{\beta}}_{\mathcal{A}_1}(0) &= argmin_{\beta_{\mathcal{A}_1}} \left\{(\mathbf{Y} - \mathbf{X}_{\mathcal{A}_1}\boldsymbol{\beta}_{\mathcal{A}_1})^T(\mathbf{Y} - \mathbf{X}_{\mathcal{A}_1}\boldsymbol{\beta}_{\mathcal{A}_1})\right\} \qquad (5.3.6) \\ &= (\mathbf{X}_{\mathcal{A}_1}^T\mathbf{X}_{\mathcal{A}_1})^{-1}\mathbf{X}_{\mathcal{A}_1}^T\mathbf{Y}, \\ \hat{\boldsymbol{\beta}}_{\mathcal{A}'}(0) &= \mathbf{0}_{\mathcal{A}'}. \qquad (5.3.7) \end{aligned}$$

So, the PLASSO estimate of $\hat{\boldsymbol{\beta}}$ at $t = 0$ is given by $\hat{\boldsymbol{\beta}} = (\hat{\boldsymbol{\beta}}_{\mathcal{A}_1}(0)^T, \hat{\boldsymbol{\beta}}_{\mathcal{A}'}(0)^T)^T$.

5.4 PLARS algorithm for solving PLASSO problem.

In this section, we propose the PLARS algorithm to build the full solution path of PLASSO problem. By characteristic, PLARS algorithm is very similar to the LARS algorithm proposed by Efron et al. [2004].

5.4.1 PLARS Algorithm

The PLARS algorithm performs the following steps in succession.

Step 0 : **[Initialization.]** Set the initial estimates :

$$\begin{aligned}
\hat{\boldsymbol{\beta}}^{(0)}_{\mathcal{A}_1} &= (\mathbf{X}^T_{\mathcal{A}_1}\mathbf{X}_{\mathcal{A}_1})^{-1}\mathbf{X}^T_{\mathcal{A}_1}\mathbf{Y}, \\
\hat{\boldsymbol{\beta}}^{(0)}_{\mathcal{A}'} &= \mathbf{0}_{\mathcal{A}'}, \\
\hat{\boldsymbol{\mu}}^{(0)} &= \mathbf{X}_{\mathcal{A}_1}(\mathbf{X}^T_{\mathcal{A}_1}\mathbf{X}_{\mathcal{A}_1})^{-1}\mathbf{X}^T_{\mathcal{A}_1}\mathbf{Y}, \\
\hat{c}^{(0)} &= \mathbf{X}^T(\mathbf{Y} - \hat{\boldsymbol{\mu}}^{(0)}), \\
\hat{c}^{(0)}_{max} &= max_{j\in\mathcal{A}'}\{|\hat{c}^{(0)}_j|\}.
\end{aligned}$$

Also, set $\mathcal{E}_0 = \mathcal{A}_1 \cup \{j : |\hat{c}_j| = C\}$ and $k = 0$.

Step (k,1) : **[Direction.]** With the current value of $\hat{\boldsymbol{c}}^{(k)}$ and $\hat{c}^{(k)}_{max}$, define

$$s_j = \begin{cases} sign(\hat{c}^{(k)}_j) & if \quad j \in \mathcal{A}' \\ 1 & if \quad j \in \mathcal{A}_1 \end{cases},$$

and $\mathcal{X}_{\mathcal{E}_k} = [...s_j\mathbf{X}_j...]_{j\in\mathcal{E}_k}$. Let $\mathbf{0}_{\mathcal{A}_1}$ be a vector of zeros of length $|\mathcal{A}_1|$, $\mathbf{1}_{\mathcal{E}_k\cap\mathcal{A}'}$ be a vector of ones of length $|\mathcal{E}_k \cap \mathcal{A}'|$ and $\mathbf{1}_{0,\mathcal{E}_k\cap\mathcal{A}'} = (\mathbf{0}^T_{\mathcal{A}_1}, \mathbf{1}^T_{\mathcal{E}_k\cap\mathcal{A}'})^T$ and define $\zeta^{(k)} = (\mathbf{1}^T_{0,\mathcal{E}_k\cap\mathcal{A}'}(\mathcal{X}^T_{\mathcal{E}_k}\mathcal{X}_{\mathcal{E}_k})^{-1}\mathbf{1}_{0,\mathcal{E}_k\cap\mathcal{A}'})^{-1/2}$. Now compute the following.

$$\begin{aligned}
\mathbf{w}^{(k)} &= \zeta^{(k)}(\mathcal{X}^T_{\mathcal{E}_k}\mathcal{X}_{\mathcal{E}_k})^{-1}\mathbf{1}_{0,\mathcal{E}_k\cap\mathcal{A}'}, \\
\mathbf{u}^{(k)} &= \mathcal{X}_{\mathcal{E}_k}\mathbf{w}_{\mathcal{E}_k},
\end{aligned}$$

$$\boldsymbol{a} = \mathcal{X}^T \mathbf{u}^{(k)},$$

Step (k,2) : **[Step Length.]** Suppose $\forall j \in \mathcal{E}_k^c \cap \mathcal{A}'$,

$$\hat{\gamma} = \left\{ \frac{\hat{c}_{max}^{(k)} - \hat{c}_j^{(k)}}{\zeta^{(k)} - \boldsymbol{a}_j}, \frac{\hat{c}_{max}^{(k)} + \hat{c}_j^{(k)}}{\zeta^{(k)} + \boldsymbol{a}_j} \right\},$$

and $\forall j \in \mathcal{E}_k \cap \mathcal{A}'$,

$$\bar{\gamma} = \left\{ -\frac{\hat{\boldsymbol{\beta}}^{(k)}}{s_j \mathbf{w}_j^{(k)}} \right\}.$$

Compute the step size $\gamma_k = min \left\{ \hat{\gamma}, \bar{\gamma}, \frac{\hat{c}_{max}^{(k)}}{\zeta^{(k)}} \right\}$.

Step (k,3) : **[Updating.]** Update the mean $\hat{\boldsymbol{\mu}}^{(k+1)}$ and the parameter estimate $\hat{\boldsymbol{\beta}}^{(k+1)}$ to

$$\begin{aligned}
\hat{\boldsymbol{\mu}}^{(k+1)} &= \hat{\boldsymbol{\mu}}^{(k)} + \gamma_k \mathbf{u}. \\
\hat{\boldsymbol{\beta}}_j^{(k+1)} &= \hat{\boldsymbol{\beta}}_j^{(k)} + \gamma_k s_j \mathbf{w}_j, j \in \mathcal{E}_k. \\
\hat{\boldsymbol{\beta}}_j^{(k+1)} &= 0, \quad j \in \mathcal{E}_k^c.
\end{aligned}$$

Update the active set as

$$\mathcal{E}_{k+1} = \begin{cases} \mathcal{E}_k \cup \{j : \gamma_j = \hat{\gamma}_j\} & if \quad \gamma_k = \hat{\gamma}, \\ \mathcal{E}_k \backslash \{j : \hat{\boldsymbol{\beta}}^{(k+1)} = 0\} & if \quad \gamma_k = \bar{\gamma}, \\ \mathcal{E}_{k+1} & if \quad \gamma_k = \frac{\hat{c}_{max}^{(k)}}{\zeta^{(k)}}. \end{cases}$$

Step (k,5) : **[Stopping Rule.]** Calculate $\hat{\boldsymbol{c}}^{(k+1)} = \mathbf{X}^T(\mathbf{Y} - \hat{\boldsymbol{\mu}}^{(k+1)})$ and $\hat{c}_{max}^{(k+1)} = max_{j \in \mathcal{A}'} \{|\hat{c}_j^{(k+1)}|\}$. If $\hat{c}_{max}^{(k+1)} = 0$, stop. Else, set $k = k + 1$ and return to Step $(k, 1)$.

5.4.2 Some properties of PLARS.

We discuss some properties of the PLARS algorithm below. In this subsection, we are mostly concerned about the direction of mean vector chosen at each step of the

algorithm. This direction dictates how the active set changes over the PLARS path.

Recall that in the PLARS algorithm, $\mathcal{E}_k$ is the active set at step k. For $\mathcal{A}_1 \subseteq \mathcal{E}_k \subseteq \{1, ..., a_1 + a_2 + a_3\}$, we define (See Step (k,1))

$$s_j = \begin{cases} sign(\hat{c}_j^{(k)}) & if \quad j \in \mathcal{A}', \\ 1 & if \quad j \in \mathcal{A}_1, \end{cases}$$

and

$$\mathcal{X}_{\mathcal{E}_k} = (...s_j \mathbf{X}_j ....)_{j \in \mathcal{E}_k}.$$

Note that $\mathbf{X}_{\mathcal{A}_1} = \mathcal{X}_{\mathcal{A}_1}$, and thus can be used interchangeably. By definition, $\hat{\boldsymbol{\mu}}^{(k)}$ is the mean vector at step k and

$$\hat{\boldsymbol{c}}^{(k)} = (\hat{c}_1^{(k)}, .., \hat{c}_p^{(k)})^T = \mathbf{X}^T(\mathbf{Y} - \hat{\boldsymbol{\mu}}^{(k)}) \tag{5.4.1}$$

is the inner product of $\mathbf{X}$ and residual vector $\mathbf{Y} - \hat{\boldsymbol{\mu}}^{(k)}$. $\hat{\boldsymbol{c}}^{(k)}$ is also often defined as the correlation vector. $\mathfrak{u}^{(k)}$ is the direction vector of PLARS, which has been defined as

$$\mathfrak{u}^{(k)} = \zeta^{(k)} \mathcal{X}_{\mathcal{E}_k} (\mathcal{X}_{\mathcal{E}_k}^T \mathcal{X}_{\mathcal{E}_k})^{-1} \mathbf{1}_{0, \mathcal{E}_k \cap \mathcal{A}'}, \tag{5.4.2}$$

where $\zeta^{(k)} = (\mathbf{1}_{0, \mathcal{E}_k \cap \mathcal{A}'}^T (\mathcal{X}_{\mathcal{E}_k}^T \mathcal{X}_{\mathcal{E}_k})^{-1} \mathbf{1}_{0, \mathcal{E}_k \cap \mathcal{A}'})^{-1/2}$.

The next proposition describes the direction vector $\mathfrak{u}^{(k)}$.

Proposition 5.1 *At each step k, the vector $\mathfrak{u}^{(k)}$ makes equal angles with the columns of $\mathcal{X}_{\mathcal{E}_k \cap \mathcal{A}'}$ and is orthogonal to columns of $\mathcal{X}_{\mathcal{A}_1}$.*

Proof: The proof follows from the fact that

$$\begin{aligned} \mathcal{X}_{\mathcal{E}_k}^T \mathfrak{u}^{(k)} &= \zeta^{(k)} \mathcal{X}_{\mathcal{E}_k}^T \mathcal{X}_{\mathcal{E}_k} (\mathcal{X}_{\mathcal{E}_k}^T \mathcal{X}_{\mathcal{E}_k})^{-1} \mathbf{1}_{0, \mathcal{E}_k \cap \mathcal{A}'} \\ &= \zeta^{(k)} \mathbf{1}_{0, \mathcal{E}_k \cap \mathcal{A}'} \\ &= \begin{pmatrix} \mathbf{0}_{\mathcal{A}_1} \\ \zeta^{(k)} \mathbf{1}_{\mathcal{E}_k \cap \mathcal{A}'} \end{pmatrix}. \end{aligned}$$

Since $\zeta^{(k)}$ is a constant, $\mathbf{X}^T_{\mathcal{E}_k \setminus \mathcal{A}_1} \mathfrak{u}^{(k)}$ is a vector with equal entries. Also, $\mathbf{X}^T_{\mathcal{A}_1} \mathfrak{u}^{(k)} = \mathcal{X}^T_{\mathcal{A}_1} \mathfrak{u}^{(k)} = \mathbf{0}_{\mathcal{A}_1}$. Therefore, Proposition 5.1 follows. □

Equiangularity with the variables which the PLASSO selects (i.e. not already known to be in the true model) is similar to the LARS algorithm. Thus PLASSO would inherit some good properties of LARS. However, unlike LARS, $\mathfrak{u}^{(k)}$ is not equiangular to all columns of $\mathcal{X}_{\mathcal{E}_k}$. On the other hand, $\mathfrak{u}^{(k)}$ is always orthogonal to the column space of $\mathbf{X}_{\mathcal{A}_1}$. This means that PLASSO chooses variables from $\mathcal{A}'$ to explain the "residuals" from the regression of $\mathbf{Y}$ on $\mathbf{X}_{\mathcal{A}_1}$. However, PLASSO is different from the procedure where we first regress $\mathbf{Y}$ on $\mathbf{X}_{\mathcal{A}_1}$, and then use ordinary LASSO to explain the residuals with variables in $\mathcal{A}'$. This is because such a procedure will hold the projection of $\mathbf{Y}$ on the column space of $\mathbf{X}_{\mathcal{A}_1}$ fixed. PLASSO estimates $\boldsymbol{\beta}$ at each step. Thus the projection on the column space of $\mathbf{X}_{\mathcal{A}_1}$ changes in each step. In terms of prediction, PLASSO should be more optimal than the "first regression then LASSO" procedure.

We now justify our proposed method to compute the statement step length γ. Notice that the initial estimate of the mean vector $\hat{\boldsymbol{\mu}}^{(0)} = \mathbf{X}\hat{\boldsymbol{\beta}}^{(k)}(0)$, which is the orthogonal projection of $\mathbf{Y}$ on the space spanned by the columns of $\mathcal{X}_{\mathcal{A}_1}$.

At step k, for some positive γ, define $\hat{\boldsymbol{\mu}}^{(k)}(\gamma)$ as

$$\hat{\boldsymbol{\mu}}^{(k)}(\gamma) = \hat{\boldsymbol{\mu}}^{(k)}(0) + \gamma \mathfrak{u}^{(k)}. \tag{5.4.3}$$

The corresponding inner product of the residual with $\mathbf{X}$ is defined as

$$\hat{\boldsymbol{c}}^{(k)}_j(\gamma) = \mathbf{X}^T(\mathbf{Y} - \hat{\boldsymbol{\mu}}^{(k)}(\gamma)).$$

Let $\hat{c}^{(k)}_{max}(\gamma) = max_j\{|\hat{c}^{(k)}_j(\gamma)|\}$.

Notice that the use of the transformation from $\mathbf{X}_{\mathcal{E}_k}$ to $\mathcal{X}_{\mathcal{E}_k}$ allows us to ignore the sign of the correlation of selected set $\mathcal{E}_k$. In particular, we have

$$\mathcal{X}^T_{\mathcal{E}_k}(\mathbf{Y} - \hat{\boldsymbol{\mu}}^{(k)}) = \hat{c}^{(k)}_{max} \mathbf{1}_{0, \mathcal{E}_k \cap \mathcal{A}'}. \tag{5.4.4}$$

With the above notations and from Proposition 5.1, we have the following result.

Proposition 5.2 *Within each step k, $\hat{\boldsymbol{c}}_{\mathcal{E}_k\cap\mathcal{A}'}$, decreases linearly with respect to γ. In particular, we have*

$$\hat{\boldsymbol{c}}_{\mathcal{E}_k\cap\mathcal{A}'}(\gamma) = (\hat{c}^{(k)}_{max} - \gamma\zeta^{(k)})\mathbf{1}_{\mathcal{E}\cap\mathcal{A}'}.$$

Proof: By Proposition 5.1, we have

$$\begin{aligned}\mathcal{X}^T_{\mathcal{E}_k}(\mathbf{Y} - \hat{\boldsymbol{\mu}}^{(k)}(\gamma)) &= \mathcal{X}^T_{\mathcal{E}_k}(\mathbf{Y} - \hat{\boldsymbol{\mu}}^{(k)} - \gamma\mathfrak{u}^{(k)}) \\ &= \hat{c}^{(k)}_{max}\mathbf{1}_{0,\mathcal{E}_k\cap\mathcal{A}'} - \gamma\zeta^{(k)}\mathbf{1}_{0,\mathcal{E}_k\cap\mathcal{A}'}.\end{aligned} \tag{5.4.5}$$

Since $\zeta^{(k)}$ is constant, Proposition 5.2 follows. □

Proposition 5.2 is a local property of PLARS for the kth step. Since $\zeta^{(k)}$ changes at each step k, the slope of $\hat{\boldsymbol{c}}_{\mathcal{E}_k\cap\mathcal{A}'}(\gamma)$ with γ for different k would be different. Next, we show that globally, the inner product between $\mathbf{X}_{\mathcal{A}_1}$ and the current residual is zero.

Proposition 5.3 *At any step k of the PLARS algorithm,*

$$\hat{\boldsymbol{c}}^{(k)}(\gamma) = \begin{pmatrix}\mathbf{0}_{\mathcal{A}_1} \\ \hat{\boldsymbol{c}}^{(k)}_{\mathcal{A}'}(\gamma)\end{pmatrix} = \begin{pmatrix}\mathbf{0}_{\mathcal{A}_1} \\ \mathbf{X}^T_{\mathcal{A}'}(\mathbf{Y} - \hat{\boldsymbol{\mu}}^{(k)}(\gamma))\end{pmatrix}.$$

Proof: By definition, $\hat{\boldsymbol{c}}_{\mathcal{A}'}(\gamma) = \mathbf{X}^T_{\mathcal{A}'}(\mathbf{Y} - \hat{\boldsymbol{\mu}}^{(k)}(\gamma))$ at any step k. To show the other equality, notice that at step 0, $\hat{\boldsymbol{c}}^{(0)}_{\mathcal{A}_1} = \mathbf{0}_{\mathcal{A}_1}$. Now from Proposition 5.1, we know that for any step k, $\hat{\boldsymbol{\mu}}^{(k+1)} = \hat{\boldsymbol{\mu}}^{(k)} + \gamma_k\mathfrak{u}^{(k)}$, where γ_k is the amount of shrinkage chosen by the PLARS algorithm as step k. Using Proposition 5.1, since $\mathbf{X}_{\mathcal{A}_1}$ remains orthogonal to $\mathfrak{u}^{(k)}$ for any k, $\hat{\boldsymbol{c}}^{(k+1)}_{\mathcal{A}_1} = \hat{\boldsymbol{c}}^{(0)}_{\mathcal{A}_1} = \mathbf{0}_{\mathcal{A}_1}$. □

Using Proposition 5.1 and 5.3, it is clear that $\hat{\boldsymbol{c}}_{\mathcal{A}_1} = \mathbf{0}_{\mathcal{A}_1}$ in the whole solution path of PLARS algorithm. This agrees with the KKT condition for the PLASSO problem (See equation (5.1)).

Next, we show at any step k, γ cannot exceed γ^*, which is defined as

$$\gamma^* = \frac{\hat{c}^{(k)}_{max}}{\zeta^{(k)}}.$$

Proposition 5.4 *If* $\gamma^* = \frac{\hat{c}^{(k)}_{max}}{\zeta^{(k)}}$, *then*

$$\mathcal{X}^T_{\mathcal{E}_k}(\mathbf{Y} - \hat{\boldsymbol{\mu}}^{(k)}(\gamma^*)) = \mathbf{0}.$$

Proof: It is straightforward to see that

$$\mathcal{X}^T_{\mathcal{E}_k}(\mathbf{Y} - \hat{\boldsymbol{\mu}}^{(k)}(\gamma^*)) = \mathcal{X}^T_{\mathcal{E}_k}(\mathbf{Y} - \hat{\boldsymbol{\mu}}^{(k)} - \gamma^*\hat{\boldsymbol{u}}^{(k)}) = \hat{c}^{(k)}_{max}\mathbf{1}_{0,\mathcal{E}_k\cap\mathcal{A}'} - \gamma^*\zeta^{(k)}\mathbf{1}_{0,\mathcal{E}_k\cap\mathcal{A}'} = \mathbf{0}.$$

□

In other words, if $\gamma_k = \gamma^*$, the current residuals is orthogonal to the column space of $\mathbf{X}_{\mathcal{E}}$, and thus the algorithm stops.

By construction, for any j in $\mathcal{E}_k$, $\hat{c}_j(\gamma) = \hat{c}_{max}(\gamma)$. Suppose at step k, there exist $j^* \in \mathcal{A}'\backslash\mathcal{E}_k$ such that for some γ, $|\hat{c}^{(k)}_{j^*}(\gamma)| = \hat{c}^{(k)}_{max}(\gamma)$. Using equation (5.4.5) and (5.4.4) with $\boldsymbol{a} = \mathcal{X}^T\mathfrak{u}^{(k)}$, this γ would satisfy

$$|\hat{c}^{(k)}_{j^*} - \gamma a_j| = \hat{c}^{(k)}_{max} - \gamma\zeta^{(k)}. \tag{5.4.6}$$

Equation (5.4.6) has two solutions. We define

$$\hat{\gamma} = min^+_{j\notin\mathcal{A}'\cap\mathcal{E}_k}\left\{\frac{\hat{c}^{(k)}_{max} - \hat{c}^{(k)}_j}{\zeta^{(k)} - a_j}, \frac{\hat{c}^{(k)}_{max} + \hat{c}^{(k)}_j}{\zeta^{(k)} + a_j}\right\}, \tag{5.4.7}$$

where min^+ means that the minimum is taken over only positive components.

Note that within step k, for some $j \in \mathcal{E}_k$. it is possible that the parameter estimate $\hat{\beta}^{(k)}_j$ would change sign. If $j \in \mathcal{E}_k \cap \mathcal{A}'$, a change in sign would violate condition (5.1). Therefore, for all $j \in \mathcal{E}_k \cap \mathcal{A}'$, we compute

$$\bar{\gamma} = -\hat{\beta}_j/s_j w_j,$$

where $\mathbf{w} = (...w_j...)^T_{j\in\mathcal{E}_k}$ is defined in step $(k,1)$ of the PLARS algorithm. The step length in the PLARS algorithm is calculated as $\gamma_k = min\{\hat{\gamma}, \bar{\gamma}, \gamma^*\}$. That is, there are three possibilities at step $(k,2)$ of the algorithm. When $\gamma_k = \hat{\gamma}$, a variable is added to

$\mathcal{E}_k$, when $\gamma_k = \bar{\gamma}$, a variable is dropped from $\mathcal{E}_k$ and when $\gamma = \gamma^*$, PLARS algorithm stops.

Next we show the proposed PLARS algorithm indeed solves the PLASSO problem (5.3.1).

5.4.3 Equivalence of PLARS and PLASSO solution path

The main result of this section can be found in Theorem 5.1 where we show that PLARS solves the PLASSO problem. This theorem follows from Lemmas 5.2 and 5.3. Both of these lemmas follow in the lines of Efron et al. [2004](See Lemma 4-10).

Lemma 5.2 *Suppose that the PLARS algorithm has just completed step $k-1$. Suppose that $\hat{\boldsymbol{\mu}}^{(k)}$, $\hat{\boldsymbol{\beta}}^{(k)}$ and $\hat{\mathbf{c}}^{(k)}$ are the current estimates of the mean, coefficient and correlation vector. Further suppose that $\mathcal{E}_k$, γ_k, $\mathbf{w}^{(k)} = \zeta^{(k)}(\mathcal{X}_{\mathcal{E}_k}^T \mathcal{X}_{\mathcal{E}_k})^{-1}\mathbf{1}_{0,\mathcal{E}_k \cap \mathcal{A}'}$ and $\mathfrak{u}^{(k)} = \mathcal{X}_{\mathcal{E}_k}\mathbf{w}_{\mathcal{E}_k}$ are the current active set, step length, coefficient direction vector and mean direction vector respectively. The following statements hold.*

(1) For $\mathcal{E}_k \backslash \mathcal{E}_{k-1} = \{j\} \subseteq \mathcal{A}'$, we have

$$\mathbf{w}_j^{(k)} \hat{\boldsymbol{c}}_j^{(k)} \geq 0.$$

Moreover, for $\gamma < \gamma_k$,

$$\hat{\boldsymbol{c}}^{(k)}(\gamma)\hat{\boldsymbol{\beta}}^{(k)}(\gamma) \geq 0.$$

(2) Let $\mathcal{S}_{\mathcal{E}_k}$ be the extended simplex with

$$\mathcal{S}_{\mathcal{E}_k} = \left\{ \sum_{j \in \mathcal{E}} s_j \mathbf{X}_j P_j : \sum_{j \in \mathcal{E} \cap \mathcal{A}'} P_j = 1 \right\}.$$

(a) The point $\mathfrak{u}^* \in \mathcal{S}_{\mathcal{E}_k}$ *with minimum* L_2 *norm is given by*

$$\mathfrak{u}^* = \zeta^{(k)}\mathfrak{u}^{(k)} = \zeta^{(k)}\mathbf{X}_{\mathcal{E}_k}\mathbf{w}^{(k)},$$

where $||\mathfrak{u}^*_{\mathcal{E}}|| = \zeta^{(k)}$.

(b) For any $\mathcal{B} \supseteq \mathcal{E}_k$, $\zeta^{(k)} \geq (\mathbf{1}^T_{0,\mathcal{B}\cap\mathcal{A}'}(\mathcal{X}^T_{\mathcal{B}}\mathcal{X}^T_{\mathcal{B}})^{-1}\mathbf{1}_{0,\mathcal{B}\cap\mathcal{A}'})^{-1/2}$.

(c) $\zeta^{(k)} \leq 1$, *with equality holding when* $|\mathcal{E}_k \cap \mathcal{A}'| = 1$.

(3) Suppose

$$\beta(\gamma) = \hat{\beta} + \gamma d \text{ and } S(\gamma) = ||y - X\beta(\gamma)||^2.$$

Let $\hat{c} = \mathbf{X}^T(\mathbf{Y} - \mathbf{X}\hat{\beta})$. *Then it follows that*

$$\begin{aligned} S(\gamma) - S(0) &= -2\hat{c}^T d\gamma + d^T\mathbf{X}^T\mathbf{X}d\gamma^2 \\ &= -2\hat{c}^T_{\mathcal{A}'}d_{\mathcal{A}'}\gamma + d^T\mathbf{X}^T\mathbf{X}d\gamma^2. \end{aligned}$$

Moreover,

$$\begin{aligned} S'(0) &= -2\hat{c}^T d \\ &= -2\hat{c}^T_{\mathcal{A}'}d_{\mathcal{A}'}. \\ S''(0) &= d^T X^T X d\gamma^2. \end{aligned}$$

Proof: Proof of Lemma 5.2 is similar to Lemma 4-6 in Efron et al. [2004]. We discuss it in the proof of Lemma 5.3. □

The first statement of Lemma 5.2 shows that when a variable is added, the corresponding estimate of regression coefficient and the inner product have the same sign. In the second statement, $\zeta^{(k)}$ is interpreted. This result is used later in (5.4.10). In the third part, we consider the change in residual sum of squares with respect to γ. This result is applicable to both PLARS and PLASSO.

In Lemma 5.3, some of the properties of PLASSO is looked at, which is related to Lemma 7-10 in Efron et al. [2004].

Lemma 5.3 *For a given t, let $\hat{\boldsymbol{\beta}}(t)$ be the PLASSO solution of (5.3.1). Let $\mathcal{E}_t = \{j : \hat{\beta}_j(t) \neq 0\}$ and $\mathcal{T}$ be an open interval such that for any $t \in \mathcal{T}$, $\mathcal{E}_t = \mathcal{E}$. Define a $|\mathcal{E}|$ by $|\mathcal{E}|$ diagonal matrix $\Delta_S = (...s_j...)_{\mathcal{E}_t}$ such that $s_j = sign(\mathbf{X}_j^T(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}(t)))$. Further, let $t_0 = inf\{T\}$. Then it follows that*

(1) The PLASSO estimates $\hat{\boldsymbol{\beta}}(t)$ and $\hat{\boldsymbol{\mu}}(t) = \mathbf{X}\hat{\boldsymbol{\beta}}(t)$ satisfy

$$\hat{\boldsymbol{\beta}}_{\mathcal{E}}(t) = \hat{\boldsymbol{\beta}}_{\mathcal{E}}(t_0) + \Delta_S \zeta_{\mathcal{E}}(t - t_0)\omega_{\mathcal{E}} \text{ and}$$
$$\hat{\boldsymbol{\mu}}_{\mathcal{E}}(t) = \hat{\boldsymbol{\mu}}_{\mathcal{E}}(t_0) + A_{\mathcal{E}}(t - t_0)\mathfrak{u}_{\mathcal{E}},$$

where $\zeta_{\mathcal{E}} = (\mathbf{1}_{0,\mathcal{E}\cap\mathcal{A}'}^T(\mathcal{X}_{\mathcal{E}}^T\mathcal{X}_{\mathcal{E}})^{-1}\mathbf{1}_{0,\mathcal{E}\cap\mathcal{A}'})^{-1/2}$, $\mathbf{w}_{\mathcal{E}} = \zeta_{\mathcal{E}}(\mathcal{X}_{\mathcal{E}}^T\mathcal{X}_{\mathcal{E}})^{-1}\mathbf{1}_{0,\mathcal{E}_k\cap\mathcal{A}'}$ and $\mathfrak{u} = \mathcal{X}_{\mathcal{E}}\mathbf{w}_{\mathcal{E}}$.

(2) Let $\hat{\mathbf{c}}(t) = \mathbf{X}_j^T(\mathbf{Y} - \hat{\boldsymbol{\mu}}(t))$, $\hat{c}_{max}(t) = max\{|\hat{c}_j(t)|, j \in \mathcal{A}' \cap \mathcal{E}\}$. Then

(a)

$$\hat{c}_j(t) = 0 \qquad , j \in \mathcal{A}_1$$
$$\hat{c}_j(t) = \hat{c}_{max} sign(\hat{\beta}_j) \qquad , j \in \mathcal{E} \cap \mathcal{A}'$$

(b)

$$\hat{c}_j(t) = 0 \qquad j \in \mathcal{A}_1.$$
$$|\hat{c}_j(t)| = \hat{c}_{max}(t) \qquad j \in \mathcal{A}' \cap \mathcal{E}.$$
$$|\hat{c}_j(t)| \leq \hat{c}_{max}(t) \qquad j \in \mathcal{E}^c.$$

(3) Define $\boldsymbol{\beta}(\gamma)$ by $\boldsymbol{\beta}(\gamma) = \hat{\boldsymbol{\beta}}(t_0) + \gamma\mathbf{d}$ for some p-vector $\mathbf{d}$ and constant γ. Also define

$$S(\gamma) = ||\mathbf{Y} - \mathbf{X}\beta(\gamma)||^2,$$
$$T(\gamma) = \sum_{j\in\mathcal{A}'} \beta_j.(\gamma)$$

If t_0 is a breakpoint of the piecewise linear solution for PLASSO solution path, the negative slope at t_0 is bounded by $2\hat{c}_{max}$, i.e.

$$R(d) = -\frac{S'(0)}{T'(0)} \leq 2\hat{c}_{max}. \tag{5.4.8}$$

Define $\mathcal{E}_1' = \{j : \hat{\beta}_j \neq 0, j \in \mathcal{A}'\}$, $\mathcal{E}_0' = \{j : \hat{\beta}_j = 0, |\hat{c}_j| = \hat{c}_{max}, j \in \mathcal{A}'\}$ and let$\mathcal{E}_{10}' = \mathcal{E}_1' \cup \mathcal{E}_0'$, and $\mathcal{E}_2' = \mathcal{A}' \backslash \mathcal{E}_{10}'$. Taking $\Delta S = S(\gamma) - S(0)$, $\Delta T = T(\gamma) - T(0)$ and $L(d) = (d^T \mathbf{X}^T \mathbf{X} d)/(\sum_{j \in \mathcal{E}_{10}'} d_j)^2$. If $d_j = 0$ for $j \in \mathcal{E}_2'$, equality holds for (5.4.8) *and*

$$\Delta S = -2\hat{c}_{max}\Delta T + L(d)^2(\Delta T)^2.$$

Proof of Lemma 5.2 and 5.3. The proof follows from Lemma 4-10 in Efron et al. [2004]. In those lemmas, we replace $\mathbf{1}_{\mathcal{E}}$ with $\mathbf{1}_{0,\mathcal{E}\cap\mathcal{A}'}$. The rest of the proof follows mutatis mutandis. □

Lemma 5.3 shows that PLASSO shares a lot of properties of the PLARS algorithm. The first part shows that with respect to t, the estimate of $\hat{\mu}$ moves along a vector that is equiangular to the vectors in $\mathbf{X}$ and orthogonal to the column space in X. One can prove the second part actually follows from the KKT optimality condition. The third part of Lemma 5.3 can be interpreted the following way. Suppose that $\mathcal{E}$ is the current active set for PLASSO. If we take $d_{\mathcal{E}} = \Delta_S w_{\mathcal{E}}$ and $d_j = 0$ for $j \in \mathcal{E}_2'$, then

$$\mathbf{X}d/(\sum_{j \in \mathcal{E}_{10}'} d_j) = \mathbf{X}_{\mathcal{E}} d_{\mathcal{E}}/(\sum_{j \in \mathcal{E}_{10}'} d_j) = \sum_{j \in \mathcal{E}} \left[\mathbf{X}_j d_j/(\sum_{j \in \mathcal{E}_{10}'} d_j) \right]. \tag{5.4.9}$$

Equation (5.4.9) implies that $\mathbf{X}d/(\sum_{j \in \mathcal{E}_{10}'} d_j)$ is in the set $\mathcal{S}_{\mathcal{E}}$ as defined in part 2 of Lemma 5.2. Thus, for any $\mathcal{E} \cap \mathcal{A}' \subseteq \mathcal{E}_{10}'$, we get $L(d) \geq \zeta_{\mathcal{E}}$ and

$$\Delta S \geq -2\hat{c}_{max}\Delta T + \zeta_{\mathcal{E}}^2(\Delta T)^2. \tag{5.4.10}$$

The equality (5.4.10) holds if $d_{\mathcal{E}} = \Delta_S w_{\mathcal{E}}$ and $d_j = 0$ for $j \notin \mathcal{E}$. Similar to LASSO, we require ΔS as negative as possible. The last part of Lemma 5.3 also says that if $d_j = 0$ for $j \notin \mathcal{E}$, then $S'(0)$ reaches its minimum. Therefore, it is sufficient to confine the support of d to $\mathcal{E}'_{10}$.

Lemma 5.4 *Define $\mathcal{E}'_1$, $\mathcal{E}'_0$ and $\mathcal{E}'_{10}$ as in part 3 of Lemma 5.3. The PLASSO satisfies the following constraint.*

(I) $\mathcal{E}'_1 \subseteq \mathcal{E} \cap \mathcal{A}'$.

(II) $\mathcal{E} \cap \mathcal{A}' \subseteq \mathcal{E}'_{10} = \mathcal{E}'_0 \cup \mathcal{E}'_1$.

(III) $w = \zeta_{\mathcal{E}}(\mathcal{X}_{\mathcal{E}}^T \mathcal{X}_{\mathcal{E}})^{-1} \mathbf{1}_{0,\mathcal{E}_k \cap \mathcal{A}'}$ *cannot have* $sign(w_j) \neq sign(\hat{c}_j)$ *for any* $j \in \mathcal{E}'_0$.

(IV) Subject to constraint I, II and III, $\mathcal{E}_k$ *must minimize* $\zeta_{\mathcal{E}}$.

(V) $-\mathbf{X}_j^T(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}) = 0$ *for all* $j \in \mathcal{A}_1$.

Proof: Our proof closely follows that of Efron et al. [2004].

(I) $\mathcal{E}'_1 \subseteq \mathcal{E} \cap \mathcal{A}'$: For some sufficiently small γ, it follows from part 1 of Lemma 5.3 that for $j \in \mathcal{E}'_1$,

$$\hat{\boldsymbol{\beta}}_j(\gamma) = \hat{\boldsymbol{\beta}}_j + \gamma s_j w_j. \tag{5.4.11}$$

Therefore, for $j \in \mathcal{E}'_1$, $\hat{\boldsymbol{\beta}}_j(\gamma)$ is nonzero when w_j is nonzero.

(II) $\mathcal{E} \cap \mathcal{A}' \subseteq \mathcal{E}'_{10}$: This follows from part 3 of lemma 5.3, which implies that the support of the coefficient direction vector d must be confined to $\mathcal{E}'_{10}$.

(III) $w = \zeta_{\mathcal{E}}(\mathcal{X}_{\mathcal{E}}^T \mathcal{X}_{\mathcal{E}})^{-1} \mathbf{1}_{0,\mathcal{E}_k \cap \mathcal{A}'}$ cannot have $sign(w_j) \neq sign(\hat{c}_j)$ for any $j \in \mathcal{E}'_0$: This is because if the signs are different, then for sufficiently small γ, $sign(\hat{\boldsymbol{\beta}}(\gamma)_j) \neq sign(\hat{c}_j)$, which violates part 2 of Lemma 5.3.

(IV) Subject to constraint I, II and III, $\mathcal{E}_k$ must minimize $\zeta^{(k)}$. This follows from part 3 of Lemma 5.3 and the requirement that (5.4.10) must be as negative as possible.

(V) $-\mathbf{X}_j^T(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}) = 0$ for all $j \in \mathcal{A}_1$: This follows from Lemma 5.1.

□

The next theorem formally shows that the solution path of PLASSO with respect to t is exactly the same as PLARS.

Theorem 5.1 *Under the PLASSO modification, with the assumption that only one variable is dropped or added at every step, the PLARS algorithm yields all PLASSO solutions.*

Proof: By Corollary 5.3, PLARS always satisfies Constraint V. From part 1 of Lemma 5.3, it is observed that PLASSO and PLARS move in the same direction as long as $\mathcal{E} \cap \mathcal{A}' = \mathcal{E}_k \cap \mathcal{A}'$. Moreover, constraint I and II in Lemma 5.4 imply that $\mathcal{E}$ cannot change since PLASSO is not at a breakpoint.

Therefore, it suffices to show that at every breakpoint of the PLASSO, the active set is equal to that obtained in the PLARS algorithm. Recall that $\hat{\boldsymbol{\beta}}(0)$ and $\hat{\boldsymbol{\beta}}^{(0)}$ are the initial estimate of $\boldsymbol{\beta}$ in the PLASSO and PLARS. The starting active sets are equal, because $\hat{\boldsymbol{\beta}}(0) = \hat{\boldsymbol{\beta}}^{(0)}$. Suppose at step k, the active set of PLASSO is $\mathcal{E} = \mathcal{E}_k$, where $\mathcal{E}_k$ is the active set of the PLARS algorithm. We show that two active sets are equal at step $k+1$. That is, same changes occur at the same places of PLARS and PLASSO.

Case 1: $\mathcal{E}_{k+1} = \mathcal{E}_k \cup \{j^*\}$ for some $j \in \mathcal{A}_1 \cap \mathcal{E}_k^c$. So by definition, $\mathcal{E}_0' = \{j^*\}$. Further, from part 1 of Lemma 5.2, $sign(w_{j^*}) = sign(\hat{c}_{j^*})$. Thus condition (III) in Lemma 5.4 is satisfied. Note that condition (I) is always true and by construction, the active set at the breakpoint is $\mathcal{E}_k$ so (II) is satisfied. Now by (IV) in Lemma 5.4 and part 2 of Lemma 5.2, it follows that the active set of PLASSO changes to $\mathcal{E}_k \cup \{j^*\}$ at this breakpoint. So PLASSO active set becomes $\mathcal{E}_{k+1}$.

Case 2: Suppose PLARS drops variable j^* from the active set at this breakpoint. So $\hat{\beta}_{j^*} = 0$ but $|\hat{c}_{j^*}| = \hat{c}_{max}$. So at the breakpoint $\mathcal{E}_1' = \mathcal{E}_k \backslash \{j^*\}$ and $\mathcal{E}_0' = \{j^*\}$. If PLASSO does not drop j^* from the active set, then $\hat{\beta}_{j^*}\hat{c}_{j^*} < 0$. So condition (III) of Lemma 5.4 would be violated. So $\mathcal{E} = \mathcal{E}_k \backslash \{j^*\} = \mathcal{E}_{k+1}$.

By induction, Theorem 5.1 holds.

□

In the next section, we establish results on the estimation consistency of the PLASSO.

5.5 Estimation consistency for PLASSO

In the previous section, we discuss computational techniques for the proposed PLASSO method. In this section, we look at some asymptotic properties of the PLASSO estimate of $\hat{\boldsymbol{\beta}}$. We use the same setup as Knight and Fu [2000], and show that under similar condition on the growth of the Lagrange multiplier λ in (5.3.2) with the sample size n, the PLASSO parameter estimate converges in probability to the minimizer of a L_1 constrained problem with non random variables. Furthermore, with correct centering and $\sqrt{n}$ scaling, a distributional convergence result can also be proved.

Recall that the signs of the component of $\boldsymbol{\beta}$ appear in the asymptotic expression of the standard LASSO estimate of $\boldsymbol{\beta}$. It is similar in PLASSO as well. To that end, we define $psign(\boldsymbol{\beta})$ as a vector with components

$$psign(\beta_j) = \begin{cases} 1 & \beta_j > 0, j \in \mathcal{A}', \\ -1 & \beta_j < 0, j \in \mathcal{A}', \\ 0 & \beta_j = 0, j \in \mathcal{A}', \\ 0 & j \in \mathcal{A}_1 \end{cases}$$

In what follows, λ is allowed to change with n. Now we use λ_n and $\hat{\boldsymbol{\beta}}^{(n)}$ to denote λ at some n and the PLASSO estimate of $\boldsymbol{\beta}$ from a sample of size n. We start with a general result in line of Knight and Fu [2000]. Some special cases are discussed at a later stage.

The following are the analogous results of Knight and Fu [2000] for PLASSO when λ_n is of order n and $\sqrt{n}$:

Theorem 5.2 *Suppose that the conditions* (5.2.2), (5.2.3) *and* (5.2.4) *hold. Let* $\mathbf{C}$ *be nonsingular.*

(1) If $\lambda_n/n \to \lambda_0 \geq 0$

$$\hat{\boldsymbol{\beta}}^{(n)} \to argmin(V_3) \text{ in probability}$$

where

$$V_3(\mathbf{u}) = (\mathbf{u} - \boldsymbol{\beta})^T \Sigma (\mathbf{u} - \boldsymbol{\beta}) + \lambda_0 \sum_{j \in \mathcal{A}'} |u_j|.$$

(2) If $\lambda_n/\sqrt{n} \to \lambda_0 \geq 0$,

$$\sqrt{n}(\hat{\boldsymbol{\beta}}^{(n)} - \boldsymbol{\beta}) \to argmin(V_4) \text{ in distribution}$$

where

$$V_4(\mathbf{u}) = -2\mathbf{u}^T\mathcal{W} + \mathbf{u}^T\Sigma\mathbf{u} + \lambda_0 \sum_{j\in\mathcal{A}'} \left[u_j sign(\beta_j) I(\beta_j \neq 0) + |u_j| I(\beta_j = 0)\right].$$

Proof: For both statements, the proof follows the steps in Knight and Fu [2000].

(a) Consider the case when $\frac{\lambda_n}{n} \to \lambda_0$. Define

$$Z_n(\phi) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \mathbf{x}_i\phi)^2 + \frac{\lambda_n}{n}\sum_{j\in\mathcal{A}'}|\phi_j|.$$

It is straightforward to see that Z_n is convex. The rest of the proof follows from Knight and Fu [2000].

(b) When $\frac{\lambda_n}{\sqrt{n}} \to \lambda_0$. Note that $\hat{\boldsymbol{u}}$ is the minimizer of

$$\begin{aligned} V_n(\mathbf{u}) &= \sum_{i=1}^{n}(y_i - \sum_{j\in A_1\cup\mathcal{A}'} x_{ij}(\beta_j + \frac{u_j}{\sqrt{n}}))^2 + \lambda_n \sum_{j\in\mathcal{A}'}|\beta_j + \frac{u_j}{\sqrt{n}}| \\ &\quad - \sum_{i=1}^{n}(y_i - \sum_{j\in A_1\cup\mathcal{A}'} x_j\beta_j)^2 - \lambda_n \sum_{j\in\mathcal{A}_2}|\beta_j| \qquad (5.5.1) \\ &= \sum_{i=1}^{n}\left[(\epsilon_i - \sum_{j\in A_1\cup\mathcal{A}'} x_{ij}\frac{u_j}{\sqrt{n}})^2 - \epsilon_i^2\right] + \lambda_n\left[\sum_{j\in\mathcal{A}'}|\beta_j + \frac{u_j}{\sqrt{n}}| - |\beta_j|\right]. \end{aligned}$$

Under (5.2.2), (5.2.3), Knight and Fu [2000] show that

$$\sum_{i=1}^{n}\left[(\epsilon_i - \sum_{j\in A_1\cup\mathcal{A}'} x_{ij}\frac{u_j}{\sqrt{n}})^2 - \epsilon_i^2\right] = -2\mathbf{u}^T\mathbf{W} + \mathbf{u}^T\mathbf{C}\mathbf{u}.$$

For all j such that $\beta_j \neq 0$, taking the Taylor expansion for $\frac{u_j}{\sqrt{n}}$ around 0 gives us

$$|\beta_j + \frac{u_j}{\sqrt{n}}| \approx |\beta_j| + sign(\beta_j)\frac{u_j}{\sqrt{n}}$$

and for all j such that $\beta_j = 0$, we have

$$|\beta_j + \frac{u_j}{\sqrt{n}}| = |\frac{u_j}{\sqrt{n}}|.$$

Since $\frac{\lambda_n}{\sqrt{n}} \to \lambda_0$, we get

$$\lambda_n \left(\sum_{j \in \mathcal{A}'} \left\{ |\beta_j + \frac{u_j}{\sqrt{n}}| - |\beta_j| \right\} \right) \to \lambda_0 \sum_{j \in \mathcal{A}'} \left[u_j sign(\beta_j) I(\beta_j \neq 0) + |u_j| I(\beta_j = 0) \right].$$

Now, since V_n is convex, it has a unique global minimum. By the epi-convergence theorem of (Geyer [1996]), we get

$$argmin(V_n) = \sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \to argmin(V) \text{ in distribution .}$$

□

When $\lambda_n = o(\sqrt{n})$, Theorem 5.2 shows that both $\hat{\boldsymbol{\beta}}_{\mathcal{A}_1}$ and $\hat{\boldsymbol{\beta}}_{\mathcal{A}'}$ are asymptotically biased. In certain special cases, $\hat{\boldsymbol{\beta}}_{\mathcal{A}_1}$ can be unbiased, but in general that won't happen.

Notice that both V_3 and V_4 are similar to their counterparts in ordinary LASSO (see V_1 and V_2 in Theorem 2.1). The only difference is that the summation in the second and third form of V_3 and V_4 respectively are over the set $\mathcal{A}'$. This is a natural modification.

We now consider a special case when $\mathcal{A}_3$ is an empty set. However, only a subset of them are known to belong to the true model.

Theorem 5.3 *Suppose that $\mathcal{A}_3 = \phi$. Let $\lambda_n/\sqrt{n} \to \lambda_0 \geq 0$ and Σ is nonsingular, then*

$$\sqrt{n}(\hat{\boldsymbol{\beta}}^{(n)} - \boldsymbol{\beta}) \to N(-\frac{\lambda_0}{2}\Sigma^{-1} psign(\boldsymbol{\beta}), \sigma^2\Sigma^{-1}) \text{ in distribution .}$$

In particular, we have

$$\sqrt{n}(\hat{\boldsymbol{\beta}}^{(n)}_{\mathcal{A}_1}-\boldsymbol{\beta}_{\mathcal{A}_1}) \to N\left(-\frac{\lambda_0}{2}\Sigma^{-1}_{\mathcal{A}_1\mathcal{A}_1}\Sigma_{\mathcal{A}_1\mathcal{A}'}\Sigma^{-1}_{\mathcal{A}'|\mathcal{A}_1} sign(\boldsymbol{\beta}_{A'}), \sigma^2(\Sigma_{\mathcal{A}_1\mathcal{A}_1|\mathcal{A}'})^{-1}\right) \text{ in distribution },$$

$$\sqrt{n}(\hat{\boldsymbol{\beta}}^{(n)}_{\mathcal{A}'} - \boldsymbol{\beta}_{\mathcal{A}'}) \to N(-\frac{\lambda_0}{2}\Sigma^{-1}_{\mathcal{A}'\mathcal{A}'|\mathcal{A}_1} sign(\boldsymbol{\beta}_{\mathcal{A}'}), \sigma^2(\Sigma_{\mathcal{A}'\mathcal{A}'|\mathcal{A}_1})^{-1}) \text{ in distribution },$$

where

$$\Sigma_{\mathcal{A}'\mathcal{A}'|\mathcal{A}_1} = \Sigma_{\mathcal{A}'\mathcal{A}'} - \Sigma_{\mathcal{A}'\mathcal{A}_1}\Sigma^{-1}_{\mathcal{A}_1\mathcal{A}_1}\Sigma_{\mathcal{A}_1\mathcal{A}'}, \Sigma_{\mathcal{A}_1\mathcal{A}_1|\mathcal{A}'} = \Sigma_{\mathcal{A}_1\mathcal{A}_1} - \Sigma_{\mathcal{A}_1\mathcal{A}'}\Sigma^{-1}_{\mathcal{A}'\mathcal{A}'}\Sigma_{\mathcal{A}'\mathcal{A}_1}.$$

Proof: Let $\mathbf{v} = \left(\mathbf{0}^T_{\mathcal{A}_1}, sign(\boldsymbol{\beta}_{\mathcal{A}'})\right)^T$. Recall that $\hat{\boldsymbol{u}} = \sqrt{n}(\hat{\boldsymbol{\beta}}^{(n)} - \boldsymbol{\beta})$. Since $\mathcal{A}_3 = \phi$, $|u_j|I(\beta_j = 0) = 0$ and by equating the derivative of $V_4(u)$ with respect to $\mathbf{u}$ to zero, we get

$$2\sqrt{n}(\Sigma(\sqrt{n}\hat{\boldsymbol{u}}) - \mathcal{W}) + \lambda_0\sqrt{n}\mathbf{v} = 0.$$

Since Σ is invertible, we have

$$\sqrt{n}\hat{\boldsymbol{u}} = \Sigma^{-1}\mathcal{W} - \frac{\lambda_0}{2}\Sigma^{-1}\mathbf{v}.$$

Since $\mathbf{W}$ follows a normal distribution and $\mathbf{v}$ is a constant, $\hat{\boldsymbol{u}}$ follows a normal distribution. Taking the expectation and variance, we get

$$E(\Sigma^{-1}\mathcal{W} - \frac{\lambda_0}{2}\Sigma^{-1}\mathbf{v}) = -\frac{\lambda_0}{2}\Sigma^{-1}\mathbf{v}.$$

$$Var(\Sigma^{-1}\mathcal{W} - \frac{\lambda_0}{2}\Sigma^{-1}\mathbf{v}) = \Sigma^{-1}Var(\mathcal{W})\Sigma^{-1} = \sigma^2\Sigma^{-1}.$$

Therefore,

$$\sqrt{n}(\hat{\boldsymbol{\beta}}^{(n)} - \boldsymbol{\beta}) \to_d N(-\frac{\lambda_0}{2}\Sigma^{-1}\mathbf{v}^T, \sigma^2\Sigma^{-1}).$$

Note that the block-wise inversion on Σ^{-1} gives us

$$\Sigma^{-1} = \begin{pmatrix} (\Sigma_{\mathcal{A}_1\mathcal{A}_1|\mathcal{A}'})^{-1} & \Sigma^{-1}_{\mathcal{A}_1\mathcal{A}_1}\Sigma_{\mathcal{A}_1\mathcal{A}'}(\Sigma_{\mathcal{A}'\mathcal{A}'|\mathcal{A}_1})^{-1} \\ -(\Sigma_{\mathcal{A}'\mathcal{A}'|\mathcal{A}_1})^{-1}\Sigma_{\mathcal{A}'\mathcal{A}_1}\Sigma^{-1}_{\mathcal{A}_1\mathcal{A}_1} & (\Sigma_{\mathcal{A}'\mathcal{A}'|\mathcal{A}_1})^{-1} \end{pmatrix}.$$

The proof is complete by noting that $\mathbf{v}$ has only zeros in the first a_1 entries. □

Theorem 5.3 shows that when $\mathcal{A}_3$ is empty, both the bias of $\hat{\boldsymbol{\beta}}_{\mathcal{A}_1}$ and $\hat{\boldsymbol{\beta}}_{\mathcal{A}'}$ depend on $\Sigma_{\mathcal{A}'\mathcal{A}_1}$, $\Sigma^{-1}_{\mathcal{A}_1\mathcal{A}_1}$ and $\Sigma_{\mathcal{A}_1\mathcal{A}'}$. Furthermore, note that if $\mathbf{X}_{\mathcal{A}_1}$ and $\mathbf{X}_{\mathcal{A}'}$ are independent, then

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_{\mathcal{A}_1} - \boldsymbol{\beta}_{\mathcal{A}_1}) \to_d N\left(0, \sigma^2(\Sigma_{\mathcal{A}_1\mathcal{A}_1})^{-1}\right).$$

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_{\mathcal{A}'} - \boldsymbol{\beta}_{\mathcal{A}'}) \to_d N(-\frac{\lambda_0}{2}\Sigma^{-1}_{\mathcal{A}'\mathcal{A}'} sign(\boldsymbol{\beta}_{A'}), \sigma^2(\Sigma_{\mathcal{A}'\mathcal{A}'})^{-1}).$$

In other words, $\hat{\boldsymbol{\beta}}_{\mathcal{A}_1}$ is asymptotically unbiased when $X_{\mathcal{A}_1}$ and $X_{\mathcal{A}'}$ are independent annd $\mathcal{A}_3$ is empty, while $\hat{\boldsymbol{\beta}}_{\mathcal{A}'}$ is always asymptotically biased.

Now, we look at the situation whereby if the true model is

$$\mathbf{Y} = \mathbf{X}_{\mathcal{A}_1}\boldsymbol{\beta}_{\mathcal{A}_1} + \mathbf{X}_{\mathcal{A}_2}\boldsymbol{\beta}_{\mathcal{A}_2},$$

and if we use the least square estimator is based on $X_{\mathcal{A}_1}$, i.e.

$$\hat{\boldsymbol{\beta}}^{ols} = (\mathbf{X}^T_{\mathcal{A}_1}\mathbf{X}_{\mathcal{A}_1})^{-1}\mathbf{X}^T_{\mathcal{A}_1}\mathbf{Y}.$$

It can be shown that the bias of the above estimator is

$$Bias(\hat{\boldsymbol{\beta}}^{ols}) = (\mathbf{X}^T_{\mathcal{A}_1}\mathbf{X}_{\mathcal{A}_1})^{-1}\mathbf{X}^T_{\mathcal{A}_1}\mathbf{X}_{\mathcal{A}_2}\boldsymbol{\beta}_{\mathcal{A}_2}.$$

The bias above is also known as the omitted-variable bias. The next corollary shows that when $Card(\mathcal{A}') = Card(\mathcal{A}_2) = 1$, the bias of $\hat{\boldsymbol{\beta}}_{\mathcal{A}_1}$ is always less than or equals to the omitted variable bias.

Corollary 5.1 *Assume that the* $Card(\mathcal{A}') = Card(\mathcal{A}_2) = 1$. *then*

$$|Bias(\hat{\boldsymbol{\beta}}_{\mathcal{A}_1})| \leq \left|(\mathbf{X}^T_{\mathcal{A}_1}\mathbf{X}_{\mathcal{A}_1})^{-1}\mathbf{X}^T_{\mathcal{A}_1}\mathbf{X}_{\mathcal{A}_2}\boldsymbol{\beta}_{\mathcal{A}_2}\right|$$

Proof: The proof is constructed using the solution path of the PLARS algorithm, which is the same as the solution path of the PLASSO. At step 0, we have

$$\hat{\boldsymbol{\beta}}^{(0)}_{\mathcal{A}_1} = (\mathbf{X}^T_{\mathcal{A}_1}\mathbf{X}_{\mathcal{A}_1})^{-1}\mathbf{X}^T_{\mathcal{A}_1}\mathbf{Y},$$
$$\hat{\boldsymbol{\beta}}^{(0)}_{\mathcal{A}_2} = 0.$$

The bias of $\hat{\boldsymbol{\beta}}^{(0)}_{\mathcal{A}_1}$ is

$$\begin{aligned} E(\hat{\boldsymbol{\beta}}^{(0)}_{\mathcal{A}_1} - \boldsymbol{\beta}_{\mathcal{A}_1}) &= E((\mathbf{X}^T_{\mathcal{A}_1}\mathbf{X}_{\mathcal{A}_1})^{-1}\mathbf{X}^T_{\mathcal{A}_1}(\mathbf{X}_{\mathcal{A}_1}\boldsymbol{\beta}_{\mathcal{A}_1} + \mathbf{X}_{\mathcal{A}_2}\beta_{\mathcal{A}_2} + \boldsymbol{\epsilon}) - \boldsymbol{\beta}_{\mathcal{A}_1}) \\ &= (\mathbf{X}^T_{\mathcal{A}_1}\mathbf{X}_{\mathcal{A}_1})^{-1}\mathbf{X}^T_{\mathcal{A}_1}\mathbf{X}_{\mathcal{A}_2}\beta_{\mathcal{A}_2}. \end{aligned}$$

By construction, since there is no other variable, only $\hat{\boldsymbol{\beta}}_{\mathcal{A}_2}$ is added to the active set at step 0. By part 1 of Lemma 5.2, the coefficient vector $\mathbf{w}^{(0)}$ has the same sign as $\hat{\boldsymbol{c}}^{(0)}_{\mathcal{A}_2}$. Furthermore, before the PLARS algorithm stops, the sign of the correlation $\hat{\boldsymbol{c}}^{(0)}_{\mathcal{A}_2}(\gamma)$ cannot change. This implies that for any $\gamma < \gamma_k$, $\hat{\boldsymbol{\beta}}_{\mathcal{A}_2}(\gamma)$ has the same sign as $\hat{\boldsymbol{c}}^{(0)}_{\mathcal{A}_2}(\gamma)$. Note that the PLARS algorithm will stop at $\hat{\boldsymbol{c}}^{(1)} = \mathbf{X}^T(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}^{(1)}) = 0$. This implies that the PLARS solution move towards least square estimate. That is,

$$\hat{\boldsymbol{\beta}}^{(1)} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}.$$

Thus the solution path only adds $\beta_{\mathcal{A}_2}$ and move linearly towards $\hat{\boldsymbol{\beta}}^{(1)}$. In the other words, for $\lambda_n \in [0, \infty)$, there exist positive γ^* such that

$$\begin{aligned} \hat{\boldsymbol{\beta}}^{(0)}(\gamma^*) &= \hat{\boldsymbol{\beta}}^{(0)} + \gamma^*(\hat{\boldsymbol{\beta}}^{(1)} - \hat{\boldsymbol{\beta}}^{(0)}) \\ &= (1 - \gamma^*)\hat{\boldsymbol{\beta}}^{(0)} + \gamma^*\hat{\boldsymbol{\beta}}^{(1)}. \end{aligned}$$

Since taking $\gamma^* = 1$ will bring us to $\hat{\boldsymbol{\beta}}^{(1)}$, γ^* is strictly less than or equals to one. Therefore, the bias of $\hat{\boldsymbol{\beta}}^{(n)}$ is bounded by

$$\left|E\left(\hat{\boldsymbol{\beta}}^{(0)}_{\mathcal{A}_1}(\gamma^*) - \boldsymbol{\beta}_{\mathcal{A}_1}\right)\right| = \left|E\left((1 - \gamma^*)\hat{\boldsymbol{\beta}}^{(0)}_{\mathcal{A}_1} + \gamma^*\hat{\boldsymbol{\beta}}^{(1)}_{\mathcal{A}_1} - \boldsymbol{\beta}_{\mathcal{A}_1}\right)\right|$$

$$
\begin{aligned}
&= \left| E\left((1-\gamma^*)(\hat{\boldsymbol{\beta}}^{(0)}_{\mathcal{A}_1} - \boldsymbol{\beta}_{\mathcal{A}_1}) + \gamma^*(\hat{\boldsymbol{\beta}}^{(1)}_{\mathcal{A}_1} - \boldsymbol{\beta}_{\mathcal{A}_1}) \right) \right| \\
&= \left| E\left((1-\gamma^*)(\hat{\boldsymbol{\beta}}^{(0)}_{\mathcal{A}_1} - \boldsymbol{\beta}_{\mathcal{A}_1}) \right) \right| \\
&\leq \left| E\left((\hat{\boldsymbol{\beta}}^{(0)}_{\mathcal{A}_1} - \boldsymbol{\beta}_{\mathcal{A}_1}) \right) \right|.
\end{aligned}
$$

□

In the next section, we establish results on the sign consistency of the PLASSO.

5.6 Sign consistency for PLASSO

Instead of the selection consistency of PLASSO, we use Partial sign consistency. Partial sign consistency is a stronger condition and is defined as follows.

Definition 5.1 Partial sign consistency *holds for* $\hat{\boldsymbol{\beta}}$ *if* $sign(\hat{\boldsymbol{\beta}}^{(n)}_{\mathcal{A}'}) = sign(\boldsymbol{\beta}_{\mathcal{A}'})$.

Partial sign consistency ignores the sign of the coefficients in $\mathcal{A}_1$. It is possible that any model selected using PLASSO may have the estimated coefficients in $\mathcal{A}_1$ having the opposite sign with the true model. However, it can also be argued that satisfying Partial Sign consistency would imply selection consistency for $\mathcal{A}$.

5.6.1 Definitions of Sign consistency and Irrepresentable conditions for PLASSO

Definition 5.2 *The PLASSO satisfies* **Partial strong sign consistency** *if there exist* $\lambda_n = f(n)$*, independent of* $\mathbf{X}$ *and* $\mathbf{Y}$*, such that*

$$lim_{n\to\infty} P(sign(\hat{\boldsymbol{\beta}}^{(n)}_{\mathcal{A}'}(\lambda_n)) = sign(\boldsymbol{\beta}_{\mathcal{A}'})) = 1. \tag{5.6.1}$$

Definition 5.3 *The PLASSO satisfies* **Partial general sign Consistency** *if there exists* λ *such that*

$$lim_{n\to\infty} P(sign(\hat{\boldsymbol{\beta}}^{(n)}_{\mathcal{A}'}(\lambda)) = sign(\boldsymbol{\beta}_{\mathcal{A}'})) = 1. \tag{5.6.2}$$

We say that if PLASSO satisfies Partial general sign consistent, then there exists $\lambda_n = \lambda$ which may depends on the sample size or some other factors such that (5.6.2) holds. However, if Partial strong sign consistency holds, λ_n only depends on the sample size. It is trivial to see that Partial strongly sign consistency implies Partial general sign consistency.

We now define the Irrepresentable conditions for PLASSO. We require the partial sample covariance $\mathbf{C}_{32|1}$ and $\mathbf{C}_{22|1}$ to be well defined and $\mathbf{C}_{22|1}$ to be invertible. Our definitions are analogous to Zhao and Yu [2006].

Definition 5.4 *We say that the* **Partial strong Irrepresentable condition** *holds if there exists a positive constant vector* $\boldsymbol{\eta}$ *such that*

$$|\mathbf{C}_{32|1}\mathbf{C}_{22|1}^{-1} sign(\boldsymbol{\beta}_{\mathcal{A}_2})| \leq 1 - \boldsymbol{\eta}. \tag{5.6.3}$$

A weaker Irrepresentable condition can also be defined.

Definition 5.5 *We say that the* **Partial weak Irrepresentable condition** *holds if*

$$|\mathbf{C}_{32|1}\mathbf{C}_{22|1}^{-1} sign(\boldsymbol{\beta}_{\mathcal{A}_2})| \leq 1. \tag{5.6.4}$$

By definition, Partial strong Irrepresentable Condition implies Partial weak Irrepresentable condition.

For the Partial Irrepresentable condition to be verified, information on whole matrix $\mathbf{C}$ except $\mathbf{C}_{33}$ is required. This is similar to Irrepresentable conditions for the standard LASSO.

5.6.2 An alternative expression of Strong Irrepresentable condition of standard LASSO

The Irrepresentable condition of standard LASSO can be re-expressed in terms of $\mathcal{A}_1$ and $\mathcal{A}_2$. It is possible that in some cases, we choose to ignore the information on $\boldsymbol{\beta}_{\mathcal{A}_1}$

and use LASSO to shrink the whole parameter vector. The proposed representation is useful for comparing PLASSO with the standard LASSO in such cases.

Corollary 5.2 *Let* $\mathbf{G} = \mathbf{C}_{31}\mathbf{C}_{1|2}^{-1} - \mathbf{C}_{32}\mathbf{C}_{22|1}^{-1}\mathbf{C}_{21}\mathbf{C}_{11}^{-1}$. *The Strong Irrepresentable Condition for LASSO can be re-expressed as :*

$$|\mathbf{G}sign(\boldsymbol{\beta}_{\mathcal{A}_1}) + \mathbf{C}_{32|1}\mathbf{C}_{22|1}^{-1}sign(\boldsymbol{\beta}_{\mathcal{A}_2})| \leq 1 - \boldsymbol{\eta}. \tag{5.6.5}$$

Similarly, the Weak Irrepresentable Condition for LASSO can be re-expressed as :

$$|\mathbf{G}sign(\boldsymbol{\beta}_{\mathcal{A}_1}) + \mathbf{C}_{32|1}\mathbf{C}_{22|1}^{-1}sign(\boldsymbol{\beta}_{\mathcal{A}_2})| \leq 1. \tag{5.6.6}$$

Proof: The strong and weak Irrepresentable condition Zhao and Yu [2006] can be expressed as

$$|\mathbf{C}_{\mathcal{A}_3\mathcal{A}}(\mathbf{C}_{\mathcal{A}\mathcal{A}})^{-1}| \leq 1 - \boldsymbol{\eta}$$

and

$$|\mathbf{C}_{\mathcal{A}_3\mathcal{A}}(\mathbf{C}_{\mathcal{A}\mathcal{A}})^{-1}| \leq 1$$

respectively. It is clear by block-wise inversion that

$$\begin{aligned}
\mathbf{C}_{\mathcal{A}_3\mathcal{A}}(\mathbf{C}_{\mathcal{A}\mathcal{A}})^{-1} &= \begin{pmatrix} \mathbf{C}_{31} & \mathbf{C}_{32} \end{pmatrix} \begin{pmatrix} \mathbf{C}_{11} & \mathbf{C}_{12} \\ \mathbf{C}_{21} & \mathbf{C}_{22} \end{pmatrix}^{-1} \\
&= \begin{pmatrix} \mathbf{C}_{31} & \mathbf{C}_{32} \end{pmatrix} \begin{pmatrix} \mathbf{C}_{1|2}^{-1} & -\mathbf{C}_{11}^{-1}\mathbf{C}_{12}\mathbf{C}_{22|1}^{-1} \\ -\mathbf{C}_{22|1}^{-1}\mathbf{C}_{21}\mathbf{C}_{11}^{-1} & \mathbf{C}_{22|1}^{-1} \end{pmatrix} \\
&= \begin{pmatrix} \mathbf{C}_{31}\mathbf{C}_{1|2}^{-1} - \mathbf{C}_{32}\mathbf{C}_{22|1}^{-1}\mathbf{C}_{21}\mathbf{C}_{11}^{-1} & \left[\mathbf{C}_{32} - \mathbf{C}_{31}\mathbf{C}_{11}^{-1}\mathbf{C}_{12}\right]\mathbf{C}_{22|1}^{-1} \end{pmatrix}.
\end{aligned}$$

Therefore, the Strong Irrepresentable condition is equivalent to

$$|\mathbf{G}sign(\boldsymbol{\beta}_{\mathcal{A}_1}) + \mathbf{C}_{32|1}(\mathbf{C}_{22|1})^{-1}sign(\boldsymbol{\beta}_{\mathcal{A}_2})| \leq 1 - \boldsymbol{\eta}.$$

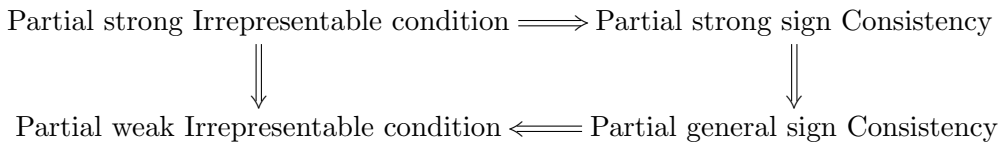


Figure 5.1 The above diagram shows the relationship berween the Partial Irrepresentable conditons and Partial sign consistency.

The Weak Irrepresentable condition follows similarly. □

Note that the second term on the left hand side of (5.6.3) and (5.6.4) is equal to the left hand side of (5.6.5) and (5.6.6) when $\mathbf{G}sign(\boldsymbol{\beta}_{\mathcal{A}_1})$ is removed. There is no restriction on the signs of the elements in $\mathbf{G}sign(\boldsymbol{\beta}_{\mathcal{A}_1})$. Thus, the Partial Irrepresentable condition is neither strong nor weaker than the Irrepresentable condition. Examples (standard regression example in Section 5.7.2 and CPG example in section 5.7.3) show that when PLASSO is consistent, LASSO may not be consistent. On the other hand, there is an example (AR(4) in section 5.7.4) where LASSO is consistent, but Partial LASSO is not.

The rest of the discussion will be split into two subsections, Partial sign consistency for finite p is presented in the section 5.6.3 while Partial sign consistency for large p is presented in the section 5.6.4.

5.6.3 Partial Sign Consistency for finite p

We have already seen that Partial sign consistency implies Partial general sign consistency and if Partial Strong Irrepresentable condition holds, Partial Weak Irrepresentable condition also holds. Under certain assumptions, for fixed p, we show that Partial Strong Irrepresentable condition implies Partial Sign consistency and Partial General Sign consistency implies Partial Weak Irrepresentable condition. Therefore, the diagram in Figure 5.1 holds.

The main result is presented in Theorem 5.4, but we require the following Lemma to prove that result (Compare [Zhao and Yu, 2006] Proposition 1).

Lemma 5.5 *Suppose there exist constant* $\boldsymbol{\eta} > 0$ *such that the Partial strong Irrepresentable condition holds. Then*

$$P\left(psign\left(\hat{\boldsymbol{\beta}}^{(n)}(\lambda_n)\right) = psign\left(\boldsymbol{\beta}\right)\right) \geq P(A_n \cap B_n)$$

where A_n *and* B_n *are defined as*

$$A_n = \left\{ |\mathbf{C}_{22|1}^{-1}\left(\mathbf{C}_{21}\mathbf{C}_{11}^{-1}\mathbf{W}_1 - \mathbf{W}_2\right)| < \sqrt{n}\left(|\boldsymbol{\beta}_{\mathcal{A}_2}| - |\mathbf{C}_{22|1}^{-1}\frac{\lambda_n}{2n} sign(\boldsymbol{\beta}_{\mathcal{A}_2})|\right)\right\}$$

$$B_n = \left\{ |\mathbf{C}_{31}\mathbf{C}_{11}^{-1}\mathbf{W}_1 - \mathbf{C}_{32|1}\mathbf{C}_{22|1}^{-1}\left(\mathbf{C}_{21}\mathbf{C}_{11}^{-1}\mathbf{W}_1 - \mathbf{W}_2\right) - \mathbf{W}_3| \leq \frac{\lambda_n}{2\sqrt{n}}\boldsymbol{\eta}\right\}$$

and

$$\mathbf{W}_1 = \frac{\mathbf{X}_{\mathcal{A}_1}^T\boldsymbol{\epsilon}}{\sqrt{n}}, \mathbf{W}_2 = \frac{\mathbf{X}_{\mathcal{A}_2}^T\boldsymbol{\epsilon}}{\sqrt{n}} \text{ and } \mathbf{W}_3 = \frac{\mathbf{X}_{\mathcal{A}_3}^T\boldsymbol{\epsilon}}{\sqrt{n}}.$$

Proof: Consider the expression $U_n(\mathfrak{u})$, which is similar to the expression used in Knight and Fu [2000] and Zhao and Yu [2006],

$$\begin{aligned} U_n(\mathfrak{u}) &= \sum_{i=1}^{n}(y_i - \sum_{j\in A_1\cup\mathcal{A}'} x_{ij}(\beta_j + u_j))^2 + \lambda_n \sum_{j\in\mathcal{A}'}|\beta_j + u_j| \\ &\quad - \sum_{i=1}^{n}(y_i - \sum_{j\in A_1\cup\mathcal{A}'} x_j\beta_j)^2 \\ &= \sum_{i=1}^{n}\left[(\epsilon_i - \sum_{j\in A_1\cup\mathcal{A}'} x_{ij}u_j)^2 - \epsilon_i^2\right] + \lambda_n\left[\sum_{j\in\mathcal{A}'}|\beta_j + u_j|\right]. \end{aligned} \tag{5.6.7}$$

It can be shown that $\hat{\boldsymbol{u}} = (\hat{\boldsymbol{\beta}}^{(n)} - \boldsymbol{\beta})$ is a minimzer of $U_n(\mathbf{u})$. Furthermore, straightforward algebra reveals that

$$\sum_{i=1}^{n}\left[(\epsilon_i - \sum_{j\in A_1\cup\mathcal{A}'} x_{ij}u_j)^2 - \epsilon_i^2\right] = -2\mathbf{W}(\sqrt{n}\mathbf{u}) + (\sqrt{n}\mathbf{u})^T\mathbf{C}(\sqrt{n}\mathbf{u})$$

and

$$\frac{d[-2\mathbf{W}(\sqrt{n}\mathbf{u}) + (\sqrt{n}\mathbf{u})^T\mathbf{C}(\sqrt{n}\mathbf{u})]}{d\mathbf{u}} = 2\sqrt{n}(\mathbf{C}(\sqrt{n}\mathbf{u}) - \mathbf{W}). \tag{5.6.8}$$

Therefore, the KKT solution of (5.6.7) must satisfy the following :

$$\begin{aligned}
&\frac{d[-2\mathbf{W}(\sqrt{n}\mathbf{u}) + (\sqrt{n}\mathbf{u})^T\mathbf{C}(\sqrt{n}\mathbf{u})]}{du_j} = 0 && \text{for } j \in \mathcal{A}_1 \\
&\frac{d[-2\mathbf{W}(\sqrt{n}\mathbf{u}) + (\sqrt{n}\mathbf{u})^T\mathbf{C}(\sqrt{n}\mathbf{u})]}{du_j} = -\lambda_n sign(u_j + \beta_j) && \text{for } j \in \mathcal{A}', \hat{\boldsymbol{u}}_j \neq 0 \\
&\frac{d[-2\mathbf{W}(\sqrt{n}\mathbf{u}) + (\sqrt{n}\mathbf{u})^T\mathbf{C}(\sqrt{n}\mathbf{u})]}{du_j} \leq \lambda_n && \text{for } j \in \mathcal{A}', \hat{\boldsymbol{u}}_j = 0.
\end{aligned}$$

Let $\hat{\boldsymbol{u}}_1$, $\hat{\boldsymbol{u}}_2$ and $\hat{\boldsymbol{u}}_3$ denote the entries corresponding to the set $\mathcal{A}_1$, $\mathcal{A}_2$ and $\mathcal{A}_3$ respectively. Consider the proposed solution

$$\sqrt{n}\mathbf{C}_{11}\hat{\boldsymbol{u}}_1 + \sqrt{n}\mathbf{C}_{12}\hat{\boldsymbol{u}}_2 - \mathbf{W}_1 = 0, \tag{5.6.9}$$

$$\sqrt{n}\mathbf{C}_{21}\hat{\boldsymbol{u}}_1 + \sqrt{n}\mathbf{C}_{22}\hat{\boldsymbol{u}}_2 - \mathbf{W}_2 = -\frac{\lambda_n}{2\sqrt{n}} sign(\boldsymbol{\beta}_{\mathcal{A}_2}), \tag{5.6.10}$$

$$\hat{\boldsymbol{u}}_3 = 0. \tag{5.6.11}$$

The proof is separated into two parts. First, it shall be shown that under condition A_n, the proposed solution gives us Partial sign consistency. That is, $sign(\hat{\boldsymbol{\beta}}^{(n)}_{\mathcal{A}_2}) = sign(\boldsymbol{\beta}_{\mathcal{A}_2})$. Second, it shall be shown that under A_n and B_n, the proposed solution is the KKT solution for (5.6.7). Because the PLASSO problem is convex, if the proposed solution is a KKT solution, then it is the global minimizer of (5.6.7).

(1) **[Under A_n, the proposed solution satisfy Partial sign consistency.]**
Rearranging equation (5.6.9), we get

$$\sqrt{n}\hat{\boldsymbol{u}}_1 = \mathbf{C}_{11}^{-1}\mathbf{W}_1 - \sqrt{n}\mathbf{C}_{11}^{-1}\mathbf{C}_{12}\hat{\boldsymbol{u}}_2. \tag{5.6.12}$$

Substituting (5.6.12) into (5.6.10), we get

$$\sqrt{n}\mathbf{C}_{21}\mathbf{C}_{11}^{-1}W_1 + \sqrt{n}\mathbf{C}_{22|1}\hat{\boldsymbol{u}}_2 - \mathbf{W}_2 = -\frac{\lambda_n}{2\sqrt{n}}sign(\boldsymbol{\beta}_{\mathcal{A}_2}). \tag{5.6.13}$$

Rearranging (5.6.13) gives us

$$\sqrt{n}\hat{\mathbf{u}}_2 = -\mathbf{C}_{22|1}^{-1}\mathbf{C}_{21}\mathbf{C}_{11}^{-1}\mathbf{W}_1 + \mathbf{C}_{22|1}^{-1}\mathbf{W}_2 - \mathbf{C}_{22|1}^{-1}\frac{\lambda_n}{2\sqrt{n}}sign(\boldsymbol{\beta}_{\mathcal{A}_2}). \tag{5.6.14}$$

Suppose that A_n is true, we have

$$\sqrt{n}|\boldsymbol{\beta}_{\mathcal{A}_2}| > |\mathbf{C}_{22|1}^{-1}\mathbf{C}_{21}\mathbf{C}_{11}^{-1}\mathbf{W}_1 - \mathbf{C}_{22|1}^{-1}\mathbf{W}_2| + |\mathbf{C}_{22|1}^{-1}\frac{\lambda_n}{2\sqrt{n}}sign(\boldsymbol{\beta}_{\mathcal{A}_2})|.$$

This implies that

$$\sqrt{n}|\boldsymbol{\beta}_{\mathcal{A}_2}| > -\mathbf{C}_{22|1}^{-1}\mathbf{C}_{21}\mathbf{C}_{11}^{-1}\mathbf{W}_1 + \mathbf{C}_{22|1}^{-1}\mathbf{W}_2 - \mathbf{C}_{22|1}^{-1}\frac{\lambda_n}{2\sqrt{n}}sign(\boldsymbol{\beta}_{\mathcal{A}_2}) \tag{5.6.15}$$

and

$$\sqrt{n}|\boldsymbol{\beta}_{\mathcal{A}_2}| > \mathbf{C}_{22|1}^{-1}\mathbf{C}_{21}\mathbf{C}_{11}^{-1}\mathbf{W}_1 - \mathbf{C}_{22|1}^{-1}\mathbf{W}_2 + \mathbf{C}_{22|1}^{-1}\frac{\lambda_n}{2\sqrt{n}}sign(\boldsymbol{\beta}_{\mathcal{A}_2}). \tag{5.6.16}$$

Therefore, from (5.6.14), (5.6.15) and (5.6.16), it is clear that

$$|\hat{\boldsymbol{u}}_2| < |\boldsymbol{\beta}_{\mathcal{A}_2}|. \tag{5.6.17}$$

The inequality constraint also can be expressed as $-|\hat{\boldsymbol{u}}_2| > -|\boldsymbol{\beta}_{\mathcal{A}_2}|$, which in turn implies that

$$sign(\boldsymbol{\beta}_{\mathcal{A}_2})\hat{\boldsymbol{u}}_2 > -|\hat{\boldsymbol{u}}_2| > -|\boldsymbol{\beta}_{\mathcal{A}_2}|. \tag{5.6.18}$$

Observe that in equation (5.6.18), for $j \in \mathcal{A}_2$, $\beta_j > 0$ implies that $\hat{\beta}_j^{(n)} - \beta_j > -\beta_j$, i.e. $\hat{\beta}_j^{(n)} > 0$, and $\beta_j < 0$ implies that $-\hat{\beta}_j^{(n)} + \beta_j > \beta_j$, i.e. $\hat{\beta}_j^{(n)} < 0$. Therefore,

it follows that the event

$$\{|\hat{\boldsymbol{u}}_2| < |\boldsymbol{\beta}_{\mathcal{A}_2}|\} \subseteq \{sign(\boldsymbol{\beta}_{\mathcal{A}_2})\hat{\boldsymbol{u}}_2 > -|\boldsymbol{\beta}_{\mathcal{A}_2}|\} \subseteq \{sign(\hat{\boldsymbol{\beta}}^{(n)}_{\mathcal{A}_2}) = sign(\boldsymbol{\beta}_{\mathcal{A}_2})\}.$$

In other words, if A_n holds, the proposed solution must have $sign(\hat{\boldsymbol{\beta}}^{(n)}_{\mathcal{A}_2}) = sign(\boldsymbol{\beta}_{\mathcal{A}_2})$. Moreover, because $\hat{\boldsymbol{u}}_3 = \hat{\boldsymbol{\beta}}^{(n)}_{\mathcal{A}_3} - \boldsymbol{\beta}_{\mathcal{A}_3} = \hat{\boldsymbol{\beta}}^{(n)}_{\mathcal{A}_3} = 0$. We have $sign(\hat{\boldsymbol{\beta}}^{(n)}_{\mathcal{A}'}) = sign(\boldsymbol{\beta}_{\mathcal{A}'})$.

(2) **[Under A_n and B_n, the proposed solution is a KKT solution.]**

For $\hat{\boldsymbol{u}}$ to be a valid KKT solution of (5.6.7), it must satisfy the following conditions.

$$\sqrt{n}\mathbf{C}_{11}\hat{\boldsymbol{u}}_1 + \sqrt{n}\mathbf{C}_{12}\hat{\boldsymbol{u}}_2 - \mathbf{W}_1 = 0, \tag{5.6.19}$$

$$\sqrt{n}\mathbf{C}_{21}\hat{\boldsymbol{u}}_1 + \sqrt{n}\mathbf{C}_{22}\hat{\boldsymbol{u}}_2 - \mathbf{W}_2 = -\frac{\lambda_n}{2\sqrt{n}} sign(\boldsymbol{\beta}_{\mathcal{A}_2} + \hat{\boldsymbol{u}}_{\mathcal{A}_2}), \tag{5.6.20}$$

$$|\sqrt{n}\mathbf{C}_{31}\hat{\boldsymbol{u}}_1 + \sqrt{n}\mathbf{C}_{32}\hat{\boldsymbol{u}}_2 - \mathbf{W}_3| \leq \frac{\lambda_n}{2\sqrt{n}}\mathbf{1}. \tag{5.6.21}$$

Therefore, first, it is required that for $j \in \mathcal{A}_2$,

$$sign(\boldsymbol{\beta}_{\mathcal{A}_2} + \hat{\boldsymbol{u}}_{\mathcal{A}_2}) = sign(\boldsymbol{\beta}_j). \tag{5.6.22}$$

Using (5.6.17), it is clear that $sign(\hat{u}_j + \beta_j) = sign(\beta_j)$, and thus equation (5.6.22) holds. Second, it is required that

$$|\sqrt{n}\mathbf{C}_{31}\hat{\boldsymbol{u}}_1 + \sqrt{n}\mathbf{C}_{32}\hat{\boldsymbol{u}}_2 - \mathbf{W}_3| \leq \frac{\lambda_n}{2\sqrt{n}}\mathbf{1}. \tag{5.6.23}$$

Now, taking the substitution of equation (5.6.12) and (5.6.14), we get

$$\begin{aligned}
&\sqrt{n}\mathbf{C}_{31}\hat{\boldsymbol{u}}_1 + \sqrt{n}\mathbf{C}_{32}\hat{\boldsymbol{u}}_2 \\
=&\mathbf{C}_{31}\mathbf{C}_{11}^{-1}\mathbf{W}_1 - \sqrt{n}\mathbf{C}_{31}\mathbf{C}_{11}^{-1}C_{12}\hat{\boldsymbol{u}}_2 + \sqrt{n}\mathbf{C}_{32}\hat{\boldsymbol{u}}_2 \qquad (5.6.24)\\
=&\mathbf{C}_{31}\mathbf{C}_{11}^{-1}\mathbf{W}_1 + \mathbf{C}_{32|1}\left[-\mathbf{C}_{22|1}^{-1}\mathbf{C}_{21}\mathbf{C}_{11}^{-1}\mathbf{W}_1 + \mathbf{C}_{22|1}^{-1}\mathbf{W}_2 - \mathbf{C}_{22|1}^{-1}\frac{\lambda_n}{2\sqrt{n}}sign(\boldsymbol{\beta}_{\mathcal{A}_2})\right].
\end{aligned}$$

Given B_n. we have

$$|\mathbf{C}_{31}\mathbf{C}_{11}^{-1}\mathbf{W}_1 - \mathbf{C}_{32|1}\mathbf{C}_{22|1}^{-1}\left(\mathbf{C}_{21}\mathbf{C}_{11}^{-1}\mathbf{W}_1 - \mathbf{W}_2\right) - \mathbf{W}_3| \leq \frac{\lambda_n}{2\sqrt{n}}\boldsymbol{\eta}.$$

Letting

$$W^* = \mathbf{C}_{31}\mathbf{C}_{11}^{-1}\mathbf{W}_1 + \mathbf{C}_{32|1}\left[-\mathbf{C}_{22|1}^{-1}\mathbf{C}_{21}\mathbf{C}_{11}^{-1}\mathbf{W}_1 + \mathbf{C}_{22|1}^{-1}\mathbf{W}_2\right] - \mathbf{W}_3,$$

we have

$$\begin{aligned}
|W^* - \mathbf{C}_{32|1}\mathbf{C}_{22|1}^{-1}\frac{\lambda_n}{2\sqrt{n}}sign(\beta_{\mathcal{A}_2})| &\leq |W^*| + |\mathbf{C}_{32|1}\mathbf{C}_{22|1}^{-1}\frac{\lambda_n}{2\sqrt{n}}sign(\beta_{\mathcal{A}_2})| \\
&\leq \frac{\lambda_n}{2\sqrt{n}}\boldsymbol{\eta} + \frac{\lambda_n}{2\sqrt{n}}(\mathbf{1} - \boldsymbol{\eta}) \\
&= \frac{\lambda_n}{2\sqrt{n}}\mathbf{1}.
\end{aligned}$$

□

We now prove the main result of this section. The following Theorem shows that under appropriate conditions, Partial strong Irrepresentable condition implies Partial strong sign consistency.

Theorem 5.4 *Suppose a_1, a_2 and a_3 are fixed and regularity Condition* (5.2.2) *and* (5.2.3) *holds. Further assume that λ_n is such that $\lambda_n/n \to 0$ and for some $0 \leq c < 1$, $\lambda_n/n^{\frac{1+c}{2}} \to \infty$. Then the Partial strong Irrepresentable condition implies*

$$P\left(psign(\hat{\boldsymbol{\beta}}^{(n)}) = psign(\boldsymbol{\beta})\right) = 1 - o(e^{-n^c}).$$

Proof: We start by defining

$$\kappa = (\kappa_{a_1+1}^T, \ldots, \kappa_{a_1+a_2}^T)^T = \mathbf{C}_{22|1}^{-1}\mathbf{C}_{21}\mathbf{C}_{11}^{-1}\frac{\mathbf{X}_{\mathcal{A}_1}^T}{\sqrt{n}} - \mathbf{C}_{22|1}^{-1}\frac{\mathbf{X}_{\mathcal{A}_2}^T}{\sqrt{n}},$$

$$\xi = (\xi_{a_1+a_2+1}^T, \ldots, \xi_{a_1+a_2+a_3}^T)^T = \left[\mathbf{C}_{31}\mathbf{C}_{11}^{-1} - \mathbf{C}_{32|1}\mathbf{C}_{22|1}^{-1}\mathbf{C}_{21}\mathbf{C}_{11}^{-1})\right]\frac{\mathbf{X}_{\mathcal{A}_1}^T}{\sqrt{n}}$$

$$+ \mathbf{C}_{32|1}\mathbf{C}_{22|1}^{-1}\frac{\mathbf{X}_{\mathcal{A}_2}^T}{\sqrt{n}} - \frac{\mathbf{X}_{\mathcal{A}_3}^T}{\sqrt{n}}.$$

Now, from the definition of A_n and B_n from the Lemma 5.5, we see that

$$\begin{aligned}
1 - P(A_n \cap B_n) \leq & P(A_n^c) + P(B_n^c) \\
& \leq \sum_{j \in \mathcal{A}_2} P\left(|\kappa_j \epsilon| \geq \sqrt{n}\left(|\beta_j| - |\frac{\lambda_n}{2n} b_j|\right)\right) \\
& \quad + \sum_{j \in \mathcal{A}_3} P(|\xi_j \epsilon| \geq \frac{\lambda_n}{2\sqrt{n}} \eta_j)
\end{aligned}$$

where $\mathbf{b} = (b_{a_1+1}, ., b_{a_1+a_2}) = \mathbf{C}_{22|1}^{-1} sign(\boldsymbol{\beta}_{\mathcal{A}_2})$.

Now, simple algebra shows that

$$\begin{aligned}
\kappa\kappa^T &= \frac{1}{n}\left[\mathbf{C}_{22|1}^{-1}\mathbf{C}_{21}\mathbf{C}_{11}^{-1}\mathbf{X}_{\mathcal{A}_1}^T - \mathbf{C}_{22|1}^{-1}\mathbf{X}_{\mathcal{A}_2}^T\right]\left[\mathbf{C}_{22|1}^{-1}\mathbf{C}_{21}\mathbf{C}_{11}^{-1}\mathbf{X}_{\mathcal{A}_1}^T - \mathbf{C}_{22|1}^{-1}\mathbf{X}_{\mathcal{A}_2}^T\right]^T \\
&= \mathbf{C}_{22|1}^{-1}\mathbf{C}_{21}\mathbf{C}_{11}^{-1}\mathbf{C}_{12}\mathbf{C}_{22|1}^{-1} - \mathbf{C}_{22|1}^{-1}\mathbf{C}_{21}\mathbf{C}_{11}^{-1}\mathbf{C}_{12}\mathbf{C}_{22|1}^{-1} - \mathbf{C}_{22|1}^{-1}\mathbf{C}_{21}\mathbf{C}_{11}^{-1}\mathbf{C}_{12}\mathbf{C}_{22|1}^{-1} \\
&\quad + \mathbf{C}_{22|1}^{-1}\mathbf{C}_{22}\mathbf{C}_{22|1}^{-1} \\
&= \mathbf{C}_{22|1}^{-1}.
\end{aligned}$$

From the regularity conditions (5.2.3), this implies

$$\kappa\epsilon \rightarrow_d N(\mathbf{0}, \Sigma_{22|1}^{-1}).$$

Similarly, it is seen that

$$\begin{aligned}
\xi\xi^T = & \left[\left(\mathbf{C}_{31}(\mathbf{C}_{11})^{-1} - (\mathbf{C}_{32|1})(\mathbf{C}_{22|1})^{-1}\mathbf{C}_{21}\mathbf{C}_{11}^{-1}\right)\frac{\mathbf{X}_1^T}{\sqrt{n}} + \mathbf{C}_{32|1}\mathbf{C}_{22|1}^{-1}\frac{\mathbf{X}_2^T}{\sqrt{n}} - \frac{\mathbf{X}_3^T}{\sqrt{n}}\right] \\
& \left[\left(\mathbf{C}_{31}\mathbf{C}_{11}^{-1} - \mathbf{C}_{32|1}\mathbf{C}_{22|1}^{-1}\mathbf{C}_{21}\mathbf{C}_{11}^{-1}\right)\frac{\mathbf{X}_1^T}{\sqrt{n}} + \mathbf{C}_{32|1}\mathbf{C}_{22|1}^{-1}\frac{\mathbf{X}_2^T}{\sqrt{n}} - \frac{\mathbf{X}_3^T}{\sqrt{n}}\right]^T \\
= & \mathbf{C}_{32|1}\mathbf{C}_{22|1}^{-1}\mathbf{C}_{21}\mathbf{C}_{11}^{-1}\mathbf{C}_{13} - \mathbf{C}_{32|1}\mathbf{C}_{22|1}^{-1}\mathbf{C}_{21}\mathbf{C}_{11}^{-1}\mathbf{C}_{12}\mathbf{C}_{22|1}^{-1}\mathbf{C}_{23|1}
\end{aligned}$$

$$
\begin{aligned}
&+ \mathbf{C}_{32|1}\mathbf{C}_{22|1}^{-1}\mathbf{C}_{22}\mathbf{C}_{22|1}^{-1}\mathbf{C}_{32|1} - \mathbf{C}_{32|1}\mathbf{C}_{22|1}^{-1}\mathbf{C}_{23} \\
&- \mathbf{C}_{31}(\mathbf{C}_{11})^{-1}\mathbf{C}_{13} - \mathbf{C}_{31}\mathbf{C}_{11}^{-1}\mathbf{C}_{12}\mathbf{C}_{22|1}^{-1}\mathbf{C}_{23|1} - \mathbf{C}_{32}\mathbf{C}_{22|1}^{-1}\mathbf{C}_{23|1} + \mathbf{C}_{33} \\
&= \mathbf{C}_{33|1} - (\mathbf{C}_{32} - \mathbf{C}_{31}\mathbf{C}_{11}^{-1}\mathbf{C}_{12})\mathbf{C}_{22|1}^{-1}\mathbf{C}_{23} \\
&= \mathbf{C}_{33|1} - \mathbf{C}_{32|1}\mathbf{C}_{22|1}^{-1}\mathbf{C}_{23|1} = \mathbf{C}_{33|12}.
\end{aligned}
$$

Thus, from the same regularity condition

$$\xi\epsilon \to_d N\left(\mathbf{0}, \Sigma_{33|1} - \Sigma_{32|1}\Sigma_{22|1}^{-1}\Sigma_{23|1}\right).$$

Now, for any $j \in \mathcal{A}_2$, since $\lambda_n/n \to 0$, and β_j is strictly greater than zero, for sufficiently large n, $\frac{\lambda_n}{n}b_j$ is much smaller than $|\beta_j|$. That is, for each $j \in \mathcal{A}_2$, $\frac{\lambda_n}{\sqrt{n}}b_j = o(1)\sqrt{n}|\beta_j|$ holds. This implies that

$$\sqrt{n}|\beta_j| - \frac{\lambda_n}{\sqrt{n}}b_j = [1 + o(1)]\sqrt{n}|\beta_j|. \tag{5.6.25}$$

Since $\kappa_j\epsilon$ and $\xi_j\epsilon$ converge in distribution to Gaussian random variables with finite variance, There exist s such that $E[(\kappa_j\epsilon)^2] = s_{\kappa_j}^2 \leq s^2$ and $E[(\xi_j\epsilon)^2] \leq s_{\xi_j}^2 \leq s^2$ for all j. The rest of the proof follows similarly to the steps used in proof of Theorem 1 in Zhao and Yu [2006]. For sufficiently large n, we get

$$\sum_{j\in\mathcal{A}_2} P\left(|\kappa_j\epsilon| \geq \sqrt{n}(|\beta_j| - \frac{\lambda_n}{2n}b_j)\right) \leq [2 + o(1)] \sum_{j\in\mathcal{A}_2}\left\{1 - \mathbf{\Phi}\left(\frac{\sqrt{n}}{s}(|\beta_j| - \frac{\lambda_n}{2n}b_j)\right)\right\}$$

where the $o(1)$ in front of the summation is to account for the approximation done on $\kappa_j\epsilon$. (Note that $\kappa_j\epsilon$ converges to Gaussian and may not be Gaussian itself). Using Mill's inequality

$$1 - \mathbf{\Phi}(t) < t^{-1}e^{-\frac{1}{2}t^2}, \tag{5.6.26}$$

we get

$$\sum_{j\in\mathcal{A}_2} P\left(|\kappa_j\epsilon| \geq \sqrt{n}(|\beta_j| - \frac{\lambda_n}{2n}b_j)\right) < [2+o(1)]a_2\left(\frac{\sqrt{n}}{s}\left(|\beta_j| - \frac{\lambda_n}{2n}b_j\right)\right)^{-1} e^{-\frac{1}{2}\left(\frac{\sqrt{n}}{s}\left(|\beta_j| - \frac{\lambda_n}{2n}b_j\right)^2\right)}. \tag{5.6.27}$$

By (5.6.25), it is clear that $\frac{\sqrt{n}}{s}(|\beta_j| - \frac{\lambda_n}{2n}b_j)$ is dominated by $\sqrt{n}$ and the RHS of (5.6.27) is dominated by e^{-n}. Since $e^{-n} = o(e^{-n^c})$ for any $0 \leq c < 1$, and a_2 is fixed, we get

$$\sum_{j\in\mathcal{A}_2} P(|\kappa_j\epsilon| \geq \sqrt{n}(|\beta_j| - \frac{\lambda_n}{2n}b_j)) = o(e^{-n^c}).$$

Now, since $n^c = o(\frac{\lambda_n^2}{n})$, using (5.6.26), we have

$$\begin{aligned}
&\sum_{j\in\mathcal{A}_3} P\left(|\xi_j\epsilon| \geq \frac{\lambda_n}{2\sqrt{n}}\eta_j\right) \\
&\leq \sum_{j\in\mathcal{A}_3} P\left(\frac{\xi_j\epsilon}{s_{\xi_j}} \geq \frac{1}{s}\frac{\lambda_n}{2\sqrt{n}}\eta_j\right) \\
&\leq (2+o(1)) \sum_{j\in\mathcal{A}_3}\left[1 - \mathbf{\Phi}\left(\frac{1}{s}\frac{\lambda_n}{2\sqrt{n}}\eta_j\right)\right] \\
&< (2+o(1)) \sum_{j\in\mathcal{A}_3}\left[\frac{1}{s}\frac{\lambda_n}{2\sqrt{n}}\eta_j\right]^{-1} e^{-\frac{1}{2s^2}\frac{\lambda_n^2}{4n}\eta_j} \\
&= o(e^{-n^c}).
\end{aligned}$$

Thus, Theorem 5.4 follows. □

Theorem 5.4 states that if there exist $\boldsymbol{\eta}$ such that Partial Strong Irrepresentable condition holds, then with high probability, there exist λ_n only depending on n such that the sign of $\hat{\boldsymbol{\beta}}^{(n)}$ is equals to $\boldsymbol{\beta}$. Our assumptions on λ_n are same as Zhao and Yu [2006] but our Irrepresentable condition is different from them. The next Theorem shows that the Partial weak Irrepresentable condition is implied by the Partial General Sign consistency.

Theorem 5.5 *Assume that a_1, a_2 and a_3 is fixed. Under regularity Condition* (5.2.2)

and (5.2.3)*, the Partial weak Irrepresentable condition is necessary for Partial general sign consistency.*

Proof: The proof is very similar to Zhao and Yu [2006]. Consider the event

$$F_1 = \{\text{There exists nonnegative } \lambda_n \text{ such that } psign(\hat{\boldsymbol{\beta}}^{(n)}(\lambda_n)) = psign(\boldsymbol{\beta})\}$$

Suppose that Partial general sign consistency holds, then

$$P(F_1) \to 1, \quad as \quad n \to \infty$$

The corresponding KKT solution (See (5.6.19), (5.6.20) and (5.6.21)) is

$$\sqrt{n}\mathbf{C}_{11}\hat{\boldsymbol{u}}_1 + \sqrt{n}\mathbf{C}_{12}\hat{\boldsymbol{u}}_2 - \mathbf{W}_1 = 0, \tag{5.6.28}$$

$$\sqrt{n}\mathbf{C}_{21}\hat{\boldsymbol{u}}_1 + \sqrt{n}\mathbf{C}_{22}\hat{\boldsymbol{u}}_2 - \mathbf{W}_2 = -\frac{\lambda_n}{2\sqrt{n}} sign(\hat{\boldsymbol{\beta}}^{(n)}_{\mathcal{A}_2}) = sign(\boldsymbol{\beta}_{\mathcal{A}_2}), \tag{5.6.29}$$

$$|\sqrt{n}\mathbf{C}_{31}\hat{\boldsymbol{u}}_1 + \sqrt{n}\mathbf{C}_{32}\hat{\boldsymbol{u}}_2 - \mathbf{W}_3| \leq \frac{\lambda_n}{2\sqrt{n}}\mathbf{1}. \tag{5.6.30}$$

Substitution of $\hat{\boldsymbol{u}}_1$ and $\hat{\boldsymbol{u}}_2$ in inequality 5.6.30 gives us (See (5.6.24))

$$|\mathbf{C}_{31}\mathbf{C}_{11}^{-1}\mathbf{W}_1 + \mathbf{C}_{32|1}\left[-\mathbf{C}_{22|1}^{-1}\mathbf{C}_{21}\mathbf{C}_{11}^{-1}\mathbf{W}_1 + \mathbf{C}_{22|1}^{-1}\mathbf{W}_2\right] - \mathbf{W}_3 - \mathbf{C}_{32|1}(\mathbf{C}_{22|1})^{-1} sign(\boldsymbol{\beta}_{\mathcal{A}_2})| \leq \frac{\lambda_n}{2\sqrt{n}}\mathbf{1}.$$

Let $W^* = \mathbf{C}_{31}\mathbf{C}_{11}^{-1}\mathbf{W}_1 + \mathbf{C}_{32|1}\left[-\mathbf{C}_{22|1}^{-1}\mathbf{C}_{21}\mathbf{C}_{11}^{-1}\mathbf{W}_1 + \mathbf{C}_{22|1}^{-1}\mathbf{W}_2\right] - \mathbf{W}_3$. Define F_2 as follows :

$$F_2 = \left\{\frac{\lambda_n}{2\sqrt{n}}L \leq W^* \leq \frac{\lambda_n}{2\sqrt{n}}U\right\}$$

$$L = -1 + \mathbf{C}_{32|1}(\mathbf{C}_{22|1})^{-1} sign(\boldsymbol{\beta}_{\mathcal{A}_2})$$

$$U = 1 + \mathbf{C}_{32|1}(\mathbf{C}_{22|1})^{-1} sign(\boldsymbol{\beta}_{\mathcal{A}_2})$$

It is clear that $F_1 \subseteq \mathcal{F}_2$, Theorem 5.5 follows from the proof of Theorem 2 in Zhao and Yu [2006]. □

Using the Theorem 5.4 and 5.5, we can see that the diagram in Figure 5.1 holds.

5.6.4 Partial Sign Consistency for Large p

When the dimension p is allowed to grow with sample size n, we shall show that Partial strong Irrepresentable condition implies the Partial strong sign consistency. However, it is not clear if whether Partial general sign consistency implies Partial weak Irrepresentable condition. Thus, the diagram in Figure 5.1 may not hold.

Also, the regularity conditions of (5.2.2) and (5.2.3) does not hold. This is because the matrix $\mathbf{X}^T\mathbf{X}/n$ may not converge. In order to control the behavior of $\mathbf{C}$ and $\boldsymbol{\beta}$, we need to make a few more assumptions. In particular, we assume that there exists constants $0 \leq c_1 < c_2 \leq 1$ and $M_1, M_2, M_3 > 0$ such that the following hold :

$$\frac{1}{n}(\mathbf{X}_i)^T\mathbf{X}_i \leq M_1, \ \ for\ all\ i, \tag{5.6.31}$$

$$\alpha^T\mathbf{C}_{22|1}\alpha \geq M_2, \ \ \forall||\alpha||_2^2 = 1, \tag{5.6.32}$$

$$a_2 = O(n^{c_1}), \tag{5.6.33}$$

$$\min_{t\in\mathcal{A}_2}\sqrt{n}|\beta_t| \geq n^{\frac{c_2}{2}}M_3. \tag{5.6.34}$$

$$a_1 + a_2 < n. \tag{5.6.35}$$

Condition (5.6.31) is required in both LASSO and PLASSO, and can be achieved by normalizing the covariates. Condition (5.6.32) requires the eigenvalues of $\mathbf{C}_{22|1}$ to be bounded from below. This essentially bounds the entries of the entries of $\mathbf{C}_{22|1}^{-1}$ from above. Since $\mathbf{C}_{22|1}^{-1}$ is submatrix of $\mathbf{C}_{\mathcal{A}\mathcal{A}}^{-1}$, we further note that (5.6.32) is weaker than what assumed by Zhao and Yu [2006]. This would imply that for every α such that $||\alpha||_2^2 = 1$,

$$\alpha^T\mathbf{C}_{\mathcal{A}\mathcal{A}}^{-1}\alpha \geq M_2$$

implies that

$$\alpha^T\mathbf{C}_{22|1}^{-1}\alpha \geq M_2.$$

Condition (5.6.33) controls the sparsity of a_2. However, for the PLASSO problem

to be solvable, we need a_2 to be smaller than n. This condition is explicitly stated in (5.6.35). Even though (5.6.35) looks a bit restrictive, but similar assumption are made in most practical circumstances. We make several observation about the Condition (5.6.34). First of all, since $\sqrt{n} > n^{c_2/2}$, a larger n would imply a relatively smaller lower bound of the minimum of $|\beta_t|$. Second, since

$$\frac{\min_{t\in\mathcal{A}_2}|\beta_t|}{n^{-1/2}} \geq n^{\frac{c_2}{2}} M_3$$

and the bias of noise terms is $O_p(n^{-1/2})$, condition (5.6.34) assumes that the minimum magnitude of $|\beta_t|$ must be at least $M_3 n^{c_2/2}$ times larger error fluctuations. This is also observed in Zhao and Yu [2006].

Last but not the least, we observe that from condition (5.6.33) and (5.6.34), we get

$$\sqrt{a_2} \leq K_1 n^{c_1/2} \leq K_1 n^{c_2/2} \leq \frac{K_1}{M_3} \min_{t\in\mathcal{A}_2} \sqrt{n}|\beta_t| \tag{5.6.36}$$

for some constant K_1. This implies that $\sqrt{a_2}$ must grow slower than $\min_{t\in\mathcal{A}_2}\sqrt{n}|\beta_t|$.

Theorem 5.6 *Assuming that ϵ_i are i.i.d random variables with at least one finite moment of order $2k$, that is for some k, $E(\epsilon)^{2k} < \infty$. Assume that the conditions (5.6.31), (5.6.32), (5.6.33), (5.6.34) and* (5.6.35) *hold. For all λ_n such that $\frac{\lambda_n}{\sqrt{n}} = o(n^{\frac{c_2-c_1}{2}})$ and $\frac{1}{a_3}\left(\frac{\lambda_n}{\sqrt{n}}\right)^{2k} \to \infty$, and $a_3 = o(n^{k(c_2-c_1)})$ holds, the Partial strong Irrepresentable condition implies that*

$$P(sign(\hat{\boldsymbol{\beta}}^{(n)}_{\mathcal{A}'}) = sign(\boldsymbol{\beta}_{\mathcal{A}'})) \geq 1 - O(\frac{a_3 n^k}{\lambda_n^{2k}}).$$

Proof: The proof is similar to Theorem 3 in Zhao and Yu [2006]. Recall the notations used in Theorem 5.4. Since $\kappa\kappa^T = \mathbf{C}^{-1}_{22|1}$, from (5.6.32), we get

$$||\kappa_j||_2^2 \leq \frac{1}{M_2}. \tag{5.6.37}$$

Further recall that $\xi\xi^T = \mathbf{C}_{33|1} - \mathbf{C}_{32|1}\mathbf{C}_{22|1}^{-1}\mathbf{C}_{23|1}$, $\mathbf{C}_{33|12}$. Let

$$P_{X_1} = \mathbf{X}_{\mathcal{A}_1}\left[(\mathbf{X}_{\mathcal{A}_1})^T\mathbf{X}_{\mathcal{A}_1}\right]^{-1}(\mathbf{X}_{\mathcal{A}_1})^T$$

be the projection matrix of the column space on $\mathbf{X}_{\mathcal{A}_1}$ and suppose

$$P_{[I-P_{X_1}]\mathbf{X}_{\mathcal{A}_2}} = [I - P_{X_1}]\,\mathbf{X}_{\mathcal{A}_2}\left[([I - P_{X_1}]\,\mathbf{X}_{\mathcal{A}_2})^T[I - P_{X_1}]\,\mathbf{X}_{\mathcal{A}_2}\right]([I - P_{X_1}]\,\mathbf{X}_{\mathcal{A}_2})^T.$$

Note that $I-P_{X_1}$ is a symmetric and idempotent matrix, therefore $[I - P_{X_1}] = [I - P_{X_1}]\,[I - P_{X_1}]$ and $[I - P_{X_1}] = [I - P_{X_1}]^T$. From the definition of $\mathbf{C}$, we get

$$\begin{aligned}
\mathbf{C}_{33|1} &= (\mathbf{X}_{\mathcal{A}_3})^T[I - P_{X_1}]\,[I - P_{X_1}]\,\mathbf{X}_{\mathcal{A}_3} \\
\mathbf{C}_{32|1}\mathbf{C}_{22|1}\mathbf{C}_{23|1} &= (\mathbf{X}_{\mathcal{A}_3})^T[I - P_{X_1}]\,\mathbf{X}_{\mathcal{A}_2}\left\{(\mathbf{X}_{\mathcal{A}_2})^T[I - P_{X_1}]\,\mathbf{X}_{\mathcal{A}_2}\right\}^{-1}(\mathbf{X}_{\mathcal{A}_2})^T[I - P_{X_1}]\,\mathbf{X}_{\mathcal{A}_3} \\
&= (\mathbf{X}_{\mathcal{A}_3})^T[I - P_{X_1}]\,P_{[I-P_{X_1}]\mathbf{X}_{\mathcal{A}_2}}[I - P_{X_1}]\,\mathbf{X}_{\mathcal{A}_3}.
\end{aligned}$$

Therefore,

$$\mathbf{C}_{33|12}\mathbf{C}_{33|1} - \mathbf{C}_{32|1}\mathbf{C}_{22|1}^{-1}\mathbf{C}_{23|1} = (\mathbf{X}_{\mathcal{A}_3})^T[I - P_{X_1}]\left\{I - P_{[I-P_{X_1}]\mathbf{X}_{\mathcal{A}_2}}\right\}[I - P_{X_1}]\,\mathbf{X}_{\mathcal{A}_3}.$$

Since $I - P_{X_1}$ and $I - P_{[I-P_{X_1}]\mathbf{X}_{\mathcal{A}_2}}$ are idempotent matrices, their eigenvalues are either 0 or 1. Thus the diagonals of $\mathbf{C}_{33|1} - \mathbf{C}_{32|1}\mathbf{C}_{22|1}(\mathbf{C}_{32|1})^T$ is dominated by the diagonals of $(\mathbf{X}_{\mathcal{A}_3})^T[I - P_{X_1}]\,\mathbf{X}_{\mathcal{A}_3}$, which in turn is dominated by the diagonals of $(\mathbf{X}_{\mathcal{A}_3})^T\mathbf{X}_{\mathcal{A}_3}$. Therefore, by assumption (5.6.31), we get

$$||\xi_i||_2^2 \leq M_1. \tag{5.6.38}$$

Therefore, given (5.6.37) and (5.6.38), if there exist $2k$th moment for ϵ, it trivially follows that

$$P(\kappa_j\epsilon > t) \leq P(|\kappa_j\epsilon| > t) \leq \frac{E[\kappa_j\epsilon]^{2k}}{t^{-2k}} = O(t^{-2k}).$$

Similarly, we have

$$P(\xi_j \epsilon > t) = O(t^{-2k}).$$

Now, using assumption (5.6.32) and using Cauchy-Schwarz inequality on the rows of $\mathbf{C}_{22|1}^{-1}$ with $sign(\beta_{\mathcal{A}_2})$, we have

$$|\mathbf{C}_{22|1}^{-1} \frac{\lambda_n}{2n} sign(\beta_{\mathcal{A}_2})| \leq \frac{\lambda_n}{nM_2} ||sign(\beta_{\mathcal{A}_2})||_2 = \frac{\lambda_n}{nM_2} \sqrt{a_2}. \tag{5.6.39}$$

Thus, for $\lambda/\sqrt{n} = o(n^{\frac{c_2-c_1}{2}})$, $\sqrt{a_2} = o(n^{c_1/2})$, using (5.6.34) and (5.6.39) and letting $b = (b_{a_1+1}, ., b_{a_1+a_2}) = \mathbf{C}_{22|1}^{-1} sign(\beta_{\mathcal{A}_2})$, we have

$$\begin{aligned}
&\sum_{j \in \mathcal{A}_2} P\left(|\kappa_j \epsilon| > \sqrt{n} \left(|\beta_j| - \frac{\lambda_n}{2\sqrt{n}} b_j \right) \right) \\
&\leq \sum_{j \in \mathcal{A}_2} P\left(|\kappa_j \epsilon| > \sqrt{n} \left(min_{j \in \mathcal{A}_2} |\beta_j| - \frac{\lambda_n}{2\sqrt{n}} b_j \right) \right) \\
&\leq \sum_{j \in \mathcal{A}_2} P\left(|\kappa_j \epsilon| > n^{\frac{c_2}{2}} M_3 - \frac{\lambda_n}{\sqrt{n} M_2} \sqrt{a_2} \right) \\
&\leq a_2 O\left(\left[n^{\frac{c_2}{2}} M_3 - \frac{\lambda_n}{\sqrt{n} M_2} \sqrt{a_2} \right]^{-2k} \right) \\
&= a_2 O\left(n^{-kc_2} \right) = O(n^{-k(c_2-c_1)}) = o(\frac{a_3 n^k}{\lambda_n^{2k}}).
\end{aligned}$$

The last equation follows because $n^{-k(c_2-c_1)}$ tends to 0 at a faster rate than either a_3 or $\frac{\lambda^{2k}}{n^k}$. For the other part, it is straightforward to show that

$$\sum_{j \in \mathcal{A}_3} P(|\xi_j| > \frac{\lambda_n}{2\sqrt{n}} \eta_i) = a_3 O\left(\frac{n^k}{\lambda_n^{2k}} \right) = O\left(\frac{a_3 n^k}{\lambda_n^{2k}} \right).$$

□

Theorem 5.6 also states the required speed at which λ_n and a_3 is allowed to grow for the PLASSO to be consistent at large p. Moreover, it is required that $\left(\frac{\lambda_n}{\sqrt{n}}\right)^{2k}$ grows to infinity faster than a_3. Also, the speed at which a_3 is allowed to grow depends on k. If only the 2nd moments of ϵ exist, then the growth of a_3 must be dominated by $n^{c_2-c_1}$,

which is smaller than n. If all 2k-th moment of ϵ exist, then a_3 would be allowed to grow at any rate. It is also clear that the growth of λ_n must be dominated by $n^{\frac{1+c_2-c_1}{2}}$, which in turn is dominated by n.

5.7 Application of PLASSO on some standard models

5.7.1 Application of PLASSO on some standard models

In this section, we inspect the difference between the selection consistency conditions between LASSO and PLASSO, with a detailed simulation study. Three models are considered. Assuming various true sub models are known, we inspect the performance of the PLASSO algorithm in selecting the correct model. This is compared with the performance of standard LASSO, where the sub model information is not used.

A partial goal of the exercise is to validate the sign consistency conditions deduced in section 5.6. This is done as follows. For each simulated dataset, we run the PLASSO and standard LASSO. We calculate the proportion of simulated data the true model appears somewhere on the path. If this proportion is "close" to one, we declare that the procedure is "consistent" for that model.

We admit that our criterion of consistency is not the usual one. However, it is often used to evaluate the performance of LASSO and other related methods (See Zhao and Yu [2006]). Moreover, we found from the simulations that there is a high correlation between the results in section 5.6 and the proportion of times we find the true model on the path.

Our simulation study show that when LASSO is inconsistent, PLASSO may be consistent. More surprising is that when PLASSO is inconsistent, LASSO may be consistent. This is strange because standard LASSO uses less information than PLASSO.

Our choice of large sample size is intentional. First of all, for large sample sizes, the sample covariance matrix would be very close to the population covariance matrix. Furthermore, both PLARS and LARS path fluctuate little if the sample size is large. We start with a standard regression model considered by Zhao and Yu [2006].

5.7.2 A standard Regression example

We generate i.i.d random variables X_{i1}, X_{i2}, e_i and ϵ_i from a $N(0,1)$, where $i = 1,\ldots,n$, and $n = 10000$, with a 3rd variable x_{i3} is constructed as

$$X_{i3} = \frac{2}{3}X_{i1} + \frac{2}{3}X_{i2} + \frac{1}{3}e_i.$$

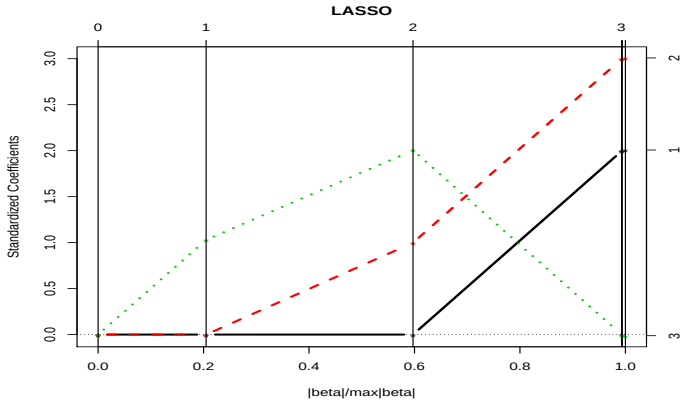
Note that X_{i3} is also i.i.d with mean 0 and variance 1 but is correlated with X_{i1} and X_{i2}. In particular, $Cov(\mathbf{X}_1, \mathbf{X}_3) = Cov(\mathbf{X}_2, \mathbf{X}_3) = 2/3$ and $Cov(\mathbf{X}_1, \mathbf{X}_2) = 0$. We generate the response vector from the equation

$$Y_i = 2X_{i1} + 3X_{i2} + \epsilon_i. \tag{5.7.1}$$

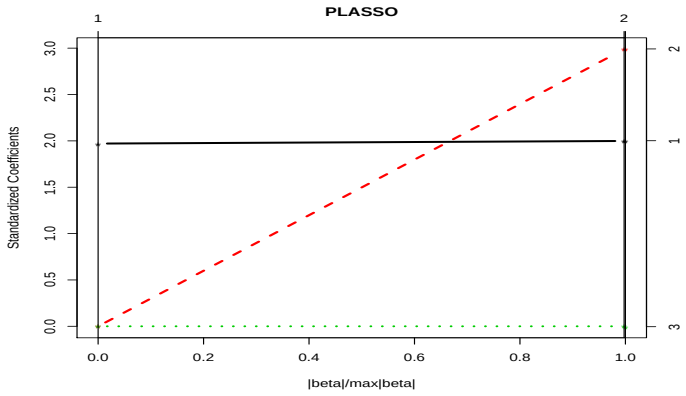
We have to recover the model from $\mathbf{X}_1$, $\mathbf{X}_2$ and $\mathbf{X}_3$.

For LASSO, the Irrepresentable condition (With $\mathbf{C}$ replaced by Σ) in Corollary 5.2 turns out to be

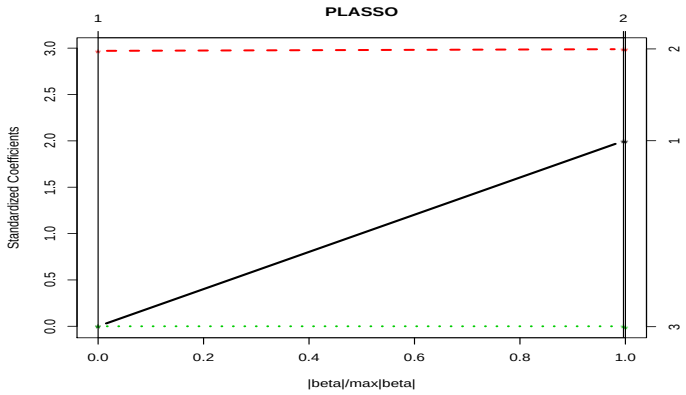
$$|\mathcal{G}sign(\boldsymbol{\beta}_{\mathcal{A}_1}) + \Sigma_{32|1}\Sigma_{22|1}^{-1}sign(\boldsymbol{\beta}_{\mathcal{A}_2})| = \frac{4}{3} > 1.$$



(a) LASSO path



(b) PLASSO Path with $\mathcal{A}_1 = \{1\}$



(c) PLASSO Path with $\mathcal{A}_1 = \{2\}$

Figure 5.2 LASSO and PLASSO path for standard regression example. The solid line represents the coefficient estimates on $\mathbf{X}_1$. The dashed line represents the coefficient estimates on $\mathbf{X}_2$. The dotted line represents the coefficient estimates on $\mathbf{X}_3$.

None of the Irrepresentable condition holds for LASSO, which means, from Zhao and Yu [2006], none of the sign consistency conditions will hold either. This is reflected in Figure 5.2(a). We can see that LASSO first select $\mathbf{X}_3$ and it never dropped in the whole path. So the LASSO procedure will never select the true model.

When we assume that $\mathcal{A}_1 = \{1\}$ and $\mathcal{A}_1 = \{2\}$, the situation changes. From Figure we have 5.2(b), it is clear that the true model is always selected. It never selects X_3. So, the correct model is always selected. And it also follows that

$$|\Sigma_{32|1}\Sigma_{22|1}^{-1} sign(\boldsymbol{\beta}_{\mathcal{A}_2})| = \frac{2}{3} < 1$$

which means both the Partial strong (5.6.3) and Partial weak Irrepresentable condition (5.6.4) are satisfied. It is actually pretty easy to show that PLARS will never pick $\mathbf{X}_3$. In this case, if we assume that $\mathcal{A}_1 = \{2\}$. then $\mathbf{X}_2^T \hat{r} = 0$. So

$$|\mathbf{X}_3^T \hat{r}| = |(\frac{2}{3}\mathbf{X}_1 + \frac{2}{3}\mathbf{X}_2 + \frac{1}{3}e)^T \hat{r}| = |(\frac{2}{3}\mathbf{X}_1 + \frac{1}{3}e)^T \hat{r}| \leq |\frac{2}{3}\mathbf{X}_1^T \hat{r}| + |\frac{1}{3}e^T \hat{r}|$$

When n is large, $|\frac{1}{3}e^T \hat{r}|$ will be significantly smaller than $|\frac{2}{3}\mathbf{X}_1^T \hat{r}|$, therefore $\mathbf{X}_1$ will always be chosen before $\mathbf{X}_3$.

5.7.3 Cocktail Party Graph(CPG) Model

Two examples of directed acyclic graph with CPG-4 and CPG-10 skeletons are presented in Figure 5.3. The directed edges are introduced in order to use these CPG models in a regression framework. The structure of the directed graph are characterized as follows.

$$pa(t) = \begin{cases} \{1, \ldots, t-1\} & \text{if } t \text{ is odd,} \\ \{1, \ldots, t-2\} & \text{if } t \text{ is even.} \end{cases} \tag{5.7.2}$$

Note that in CPG model, $p/2$ pairs of nodes are not adjacent. For example, in CPG-4, the pairs $(1, 2)$ and $(3, 4)$ are not adjaccent. In CPG-10 (see figure 5.3), these pairs $(1, 2), (3, 4), (5, 6), (7, 8)$ and $(9, 10)$ are not adjacent. We denote these pairs as partners.

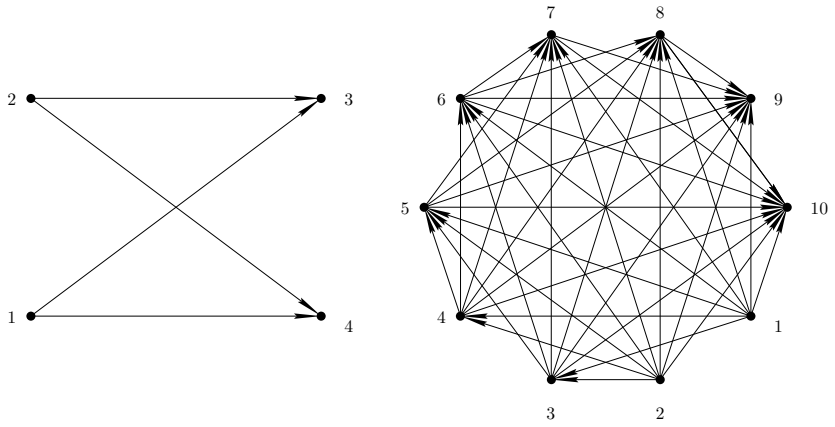


Figure 5.3 Two example of CPG model : CPG-4 and CPG-10

We try to recover the set of parents of node 4 in CPG-4 and node 10 in CPG-10. Note that in CPG-4, the parents of node 4 is 1 and 2. In CPG-10, the parents of node 10 is 1 to 8. The covariates are highly correlated because of the graph structure. The graph are parameterized as follows. Assume that for each t, node t represents a normal random variable with

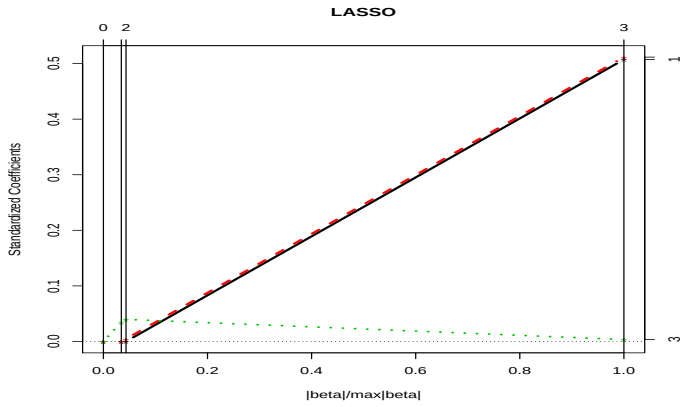
$$E[v_t|pa(v_t)] = \sum_{j \in pa(t)} 0.75 v_j,$$

and

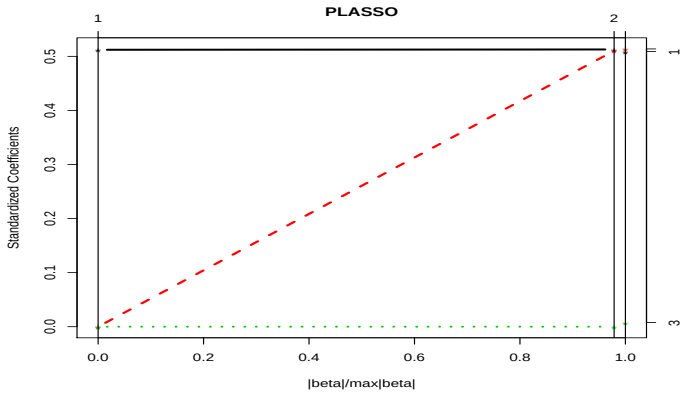
$$Var[v_t|pa(v_t)] = 1.$$

Let R be the correlation matrix of $v = (v_1, .., v_p)$. In order to maintain unit variance, we generate $\mathbf{X} \sim N(\mathbf{0}, R)$. Note that this does not violate the conditional independence relationships represented by the graphs. The sample size is $n = 10000$. The result quoted are average of 100.

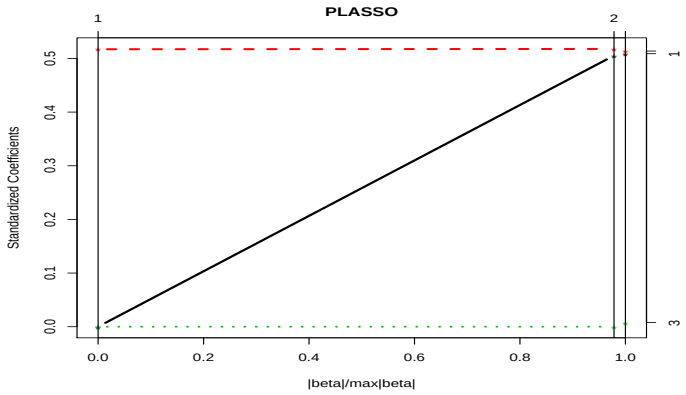
Figure 5.4 shows simulated paths of LASSO AND PLASSO for CPG-4. Notice that LASSO selects the edge $(3, 4)$ on the first step and never shrinks it back to zero subsequently. PLASSO with $\mathcal{A}_1 = \{(1, 4)\}$ or $\mathcal{A}_1 = \{(2, 4)\}$ almost always selects the correct model for almost whole of the path.



(a) LASSO path with no known variables



(b) PLASSO path with $\mathcal{A}_1 = \{(1,4)\}$ known



(c) PLASSO path with $\mathcal{A}_1 = \{(2,4)\}$ known

Figure 5.4 An example of paths for LASSO and PLASSO on CPG-4. The solid line represents the edge $(1,4)$, dashed line represents the edge $(2,4)$ while the dotted line represents the edge from $(3,4)$.

Table 5.1 Simulation results using PLASSO for CPG-10.

Known Edges	$\Sigma_{32\|1}\Sigma_{2\|1}^{-1} Sign(\boldsymbol{\beta}_{\mathcal{A}_2})$	#Inconsistent Path	First edge Added
ϕ	-1.058	51	$(9,10)$
$(1,10)$	-1.017	53	$(2,10)$
$(2,10)$	-1.017	50	$(1,10)$
$(3,10)$	-0.998	9	$(4,10)$
$(4,10)$	-0.998	6	$(3,10)$
$(5,10)$	-0.933	0	$(6,10)$
$(6,10)$	-0.933	0	$(5,10)$
$(7,10)$	-0.756	0	$(8,10)$
$(8,10)$	-0.756	0	$(7,10)$
$(1,10)(2,10)$	-0.975	0	$(3,10)/(4,10)$
$(1,10)(3,10)$	-0.956	0	$(4,10)$
$(1,10)(4,10)$	-0.956	0	$(3,10)$
$(1,10)(5,10)$	-0.891	0	$(6,10)$
$(1,10)(6,10)$	-0.891	0	$(5,10)$
$(1,10)(7,10)$	-0.715	0	$(8,10)$
$(1,10)(8,10)$	-0.715	0	$(7,10)$
$(2,10)(3,10)$	-0.956	0	$(4,10)$
$(2,10)(4,10)$	-0.956	0	$(3,10)$
$(2,10)(5,10)$	-0.891	0	$(6,10)$
$(2,10)(6,10)$	-0.891	0	$(5,10)$
$(2,10)(7,10)$	-0.715	0	$(8,10)$
$(2,10)(8,10)$	-0.715	0	$(7,10)$
$(3,10)(4,10)$	-0.937	0	$(1,10)/(2,10)$
$(3,10)(5,10)$	-0.872	0	$(6,10)$
$(3,10)(6,10)$	-0.872	0	$(5,10)$
$(3,10)(7,10)$	-0.696	0	$(8,10)$
$(3,10)(8,10)$	-0.696	0	$(7,10)$
$(4,10)(5,10)$	-0.872	0	$(6,10)$
$(4,10)(6,10)$	-0.872	0	$(5,10)$
$(4,10)(7,10)$	-0.696	0	$(8,10)$
$(4,10)(8,10)$	-0.696	0	$(7,10)$
$(5,10)(6,10)$	-0.807	0	$(3,10)/(4,10)$
$(5,10)(7,10)$	-0.631	0	$(6,10)$
$(5,10)(8,10)$	-0.631	0	$(6,10)$
$(6,10)(7,10)$	-0.631	0	$(5,10)$
$(6,10)(8,10)$	-0.631	0	$(5,10)$
$(7,10)(8,10)$	-0.454	0	$(5,10)/\ (6,10)$

In our simulation study, we observe that LASSO chooses the wrong parent, which is node 9, throughout the path and at 50% of the time drops before finishing. We conclude that the LASSO doesn't show sign consistency for this model.

Results for PLASSO can be found in Table 5.1, where we present the percentage of inconsistent path seen and the first variable added for different choices of $\mathcal{A}_1$.

For CPG-10 model, it can be seen that the LASSO Irrepresentable condition is violated because

$$\mathcal{G}sign(\boldsymbol{\beta}_{\mathcal{A}_1}) + \Sigma_{32|1}\Sigma_{2|1}^{-1}sign(\boldsymbol{\beta}_{\mathcal{A}_2}) = -1.058 < -1.$$

As observed in Table 5.1, when the known parent is node $5, 6, 7$ and 8, then PLASSO is consistent.

When the known parent is node 1 or 2, the PLASSO become sign inconsistent, which is reflected in both the Irrepresentable condition value and the porportion of times the true model is selected. When $\mathcal{A}_1 = \{(10, 3)\}$ or $\{(10, 4)\}$, we get a very low porportions of inconsistent path which is because of random flunataion. Increasing the sample size to 100,000 produces completely consistent path. When $\mathcal{A}_1 \subseteq \{(10, 5), (10, 6), (10, 7), (10, 8)\}$ or two parents are specified, all path are consistent.

Another interesting observation is that if only one of the parents of node 10 is known, the first step of PLASSO is to add the parent's paired partner. For example, if $\mathcal{A}_1 = \{(3, 10)\}$, at the first step, the edge from 4 to 10 is always selected.

5.7.4 Fourth order Autoregressive (AR(4)) Model

We consider an AR(4) model with 10 nodes (see Figure 5.5). The parameterization is same as CPG model. We want to identify the parents set of node 10. The true parents set is $\{6, 7, 8, 9\}$. In this case,

$$\mathcal{G}sign(\boldsymbol{\beta}_{\mathcal{A}_1}) + \Sigma_{32|1}\Sigma_{2|1}^{-1}sign(\boldsymbol{\beta}_{\mathcal{A}_2}) = \begin{pmatrix} -0.791 \\ -0.867 \\ -0.940 \\ -0.976 \\ -0.989 \end{pmatrix}.$$

Therefore, LASSO is sign consistent and in our simulation the true model is on the LASSO path 100% of the times. The strange thing for this model is that PLASSO is

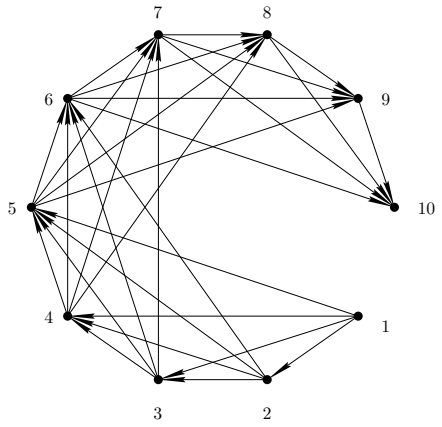


Figure 5.5 AR4 with 10 nodes

never consistent for any specification of $\mathcal{A}_1$. The results are presented in table 5.2 . We note that both Partial Strong and Weak Irrepresentable fails for at least one vertex so PLASSO wont be sign consistent in this case and the simulation confirms it.

Table 5.2 Simulation results using PLASSO for AR(4) model.

Known coefficients	$\Sigma_{32\|1}\Sigma_{2\|1}^{-1}Sign(\boldsymbol{\beta}_{\mathcal{A}_2})$					#Inconsistent Path
$(6,10)$	-2.004	0.734	-0.667	-1.214	-1.458	100
$(7,10)$	-1.395	-1.690	1.005	-0.813	-1.444	100
$(8,10)$	-0.197	-1.159	-1.800	0.922	-1.000	100
$(9,10)$	1.224	-0.487	-1.358	-1.824	0.935	100
$(6,10)(7,10)$	-2.608	-0.088	1.278	-1.050	-1.913	100
$(6,10)(8,10)$	-1.410	0.442	-1.527	0.684	-1.469	100
$(6,10)(9,10)$	0.011	1.115	-1.085	-2.061	0.466	100
$(7,10)(8,10)$	-0.802	-1.982	0.146	1.085	-1.455	100
$(7,10)(9,10)$	0.620	-1.309	0.587	-1.660	0.480	100
$(8,10)(9,10)$	1.817	-0.779	-2.218	0.074	0.924	100
$(6,10)(7,10)(8,10)$	0.620	-1.309	0.587	-1.660	0.480	100
$(6,10)(7,10)(9,10)$	-0.593	0.292	0.860	-1.898	0.011	100
$(7,10)(8,10)(9,10)$	1.213	-1.602	-0.273	0.238	0.469	91

5.8 Discussion

In this chapter, we propose a natural extension of the LASSO method called PLASSO, so that variables known to be in the true model are always picked. We also proposed

a PLARS algorithm, which is adapted from the LARS algorithm [Efron et al., 2004] to solve the PLASSO problem.

Furthermore, we also look at some selection consistency conditions for PLASSO. We show that the Partial Irrepresentable conditions for LASSO differs from the LASSO Irrepresentable conditions. In particular, Partial Irrepresentable conditions are also neither stronger or weaker.

Our theoretical results are based on linear model. However, PLASSO can be used for graphical Markov model selection, such as on DAG.

We note that PLASSO-based methods can also be used with many available methods, such as the adaptive LASSO [Zou, 2006] and elastic net [Zou and Hastie, 2005]. Our result offers insights into the benefits and problems that PLASSO may have in its future extensions.

CHAPTER 6

Almost Qualitative Comparison of Signed Partial Correlation

6.1 Introduction

Graphical models are specified by conditional independence relationships. Several algorithms to read off these conditional independences from a given graph has been postulated. The path based separation for various models like undirected graphs and directed acyclic graphs [Verma and Pearl, 1990] are known. Available completeness results ensure that if relevant connection criterion is satisfied by two vertices a and c (*correlates*) given a set of vertices Z (*conditionate*) then in almost all distributions "factoring" according to the graph, a and c would be conditionally dependent given Z.

However, it is known that all such connecting paths don't represent equal conditional dependence. In many cases shorter paths imply stronger dependence [Greenland, 2003], in many others they don't [Chaudhuri and Richardson, 2003, Theorem 2].

A more general problem is to order the conditional dependence among the components of a Gaussian random vector. For these vectors the conditional dependencies are completely specified by the partial correlation and regression coefficients. It can also be

shown that the coefficients are polynomials of the entries in the covariance matrix of the vector.

Thus in many situations it would be beneficial to be able to compare these coefficients in a way such that their ordering does not depend on the particular entries in the covariance matrix. We denote such parameter value independent ordering as *qualitative*.

Algorithms based on conditional independence relations usually require faithfulness assumption. Uhler et al. [2013], Lin et al. [2012] explore bounds of deviation from faithfulness of the graph to its underlying distribution. Therefore, qualitative comparison can be used to specify such bounds.

Simple counter examples show that such qualitative comparisons cannot hold in general. However, if the covariance matrix satisfies certain conditional independence criterion, some squared correlation coefficients can be qualitatively compared. Chaudhuri and Richardson [2003], Chaudhuri [2013] provide such results. These results can be applied to several graphical models, where the validity of the sufficient conditional independence relationships can be easily read off. Rules for comparing strength of connection on tree and polytree models are already known. Chaudhuri and Tan [2010] do the same for absolute values of partial regression coefficients.

Such measures of degree of association however does not tell the full story. For Gaussian random vectors the signs of the partial correlations are important too. Thus it is interesting to enquire if the signed partial correlation and regression coefficients can be qualitatively ordered like their squares as well.

In this chapter we address such issues. We show that qualitative comparison of signed partial correlation and regression coefficients do not hold except in some special cases. However, under certain conditions the nature of comparison is dependent only on the sign of certain partial covariances. They are not dependent on the particular values in the covariance matrix. We term such comparisons as *almost qualitative*.

We show that, the number of covariances determining the comparison is also minimal and in most cases can be determined from the observed data. We further show that, for trees and a class of polytrees they are completely determined by the vertices on the

connecting path.

We also apply our results to single factor models to obtain a necessary and sufficient characterization for them. In particular, the results give us a method to identify single factor models when information is partially observed. That is, when there is a hidden variable that is not observed.

The rest of the chapter is organized as follows. Section 6.2 contains a description of the notations and some preliminary definition used in chapter. Three key situations are discussed in Section 6.3. These situations apply to a general Gaussian random vector and are not specific to any graphical models. In Section 6.4, 6.5 and 6.6 we apply our key results to Gaussian tree and polytree models. Section 6.7 consists of the characterization of single factor models.

6.2 Notation and Initial Definitions

Suppose $V \sim N(\mu, \Sigma)$ with a positive definite Σ. Let a, b, c, c', z , z', x etc. be the components and B, Z etc. be the subsets of components of V. In this chapter V will also denote the vertex set of the underlying graph G. The conditional covariance (ie. $\sigma_{ac|Z}$) between a and c given a subset Z is given by

$$\sigma_{ac|Z} = \sigma_{ac} - \Sigma_{aZ}\Sigma_{ZZ}^{-1}\Sigma_{Zc}.$$

Here σ_{ab} and Σ_{aZ} respectively denote the (a, b)th element and $a \times Z$ submatrix of Σ. If $Z = \{z_1, z_2, \ldots, z_p\}$, then $\sigma_{ac|Z}$ can be iteratively computed as [Kendall and Stuart, 1979, Brito and Pearl, 2002],

$$\sigma_{ac|z_1 \ldots z_p} = \sigma_{ac|z_1 \ldots z_{p-1}} - \frac{\sigma_{az_p|z_1 \ldots z_{p-1}}\sigma_{cz_p|z_1 \ldots z_{p-1}}}{\sigma_{z_p z_p|z_1 \ldots z_{p-1}}}$$

The partial correlation between a and c given Z (ie. $\rho_{ac|Z}$) and the partial regression

coefficient of a on c given Z (ie. $\beta_{ac|Z}$) is defined as

$$\rho_{ac|Z} = \frac{\sigma_{ac|Z}}{\sqrt{\sigma_{aa|Z}\sigma_{cc|Z}}},\ \beta_{ac|Z} = \frac{\sigma_{ac|Z}}{\sigma_{cc|Z}}.$$

Also, throughout this chapter we keep the dependent vertices (correlates) fixed and compare their dependence by varying the set of vertices conditioned on (conditionates). Comparisons with fixed conditionate are not attempted here.

Recall that a graph is denoted as $G = (V, E)$, where V is the set of vertices and E is the set of edges. For UG and DAG, we follow the definitions and notations from Chapter 3.

In what follows, $X \propto^+ Y$ means that X and Y have the same sign or $X = M \cdot Y$ for some non-negative constant M.

Suppose $\mathcal{S}^+$ denote the set of all positive definite covariance matrices. Clearly partial correlation and regression coefficients are functions of $\Sigma \in \mathcal{S}^+$ taking values in $\mathbb{R}$. We now define qualitative and almost qualitative comparisons for two functions f and g of Σ over a subset $\mathcal{T}$ of $\mathcal{S}^+$.

Definition 6.1 *We say f is qualitatively larger (smaller) than g on $\mathcal{T} \subseteq \mathcal{S}^+$, if $f(\Sigma) > g(\Sigma)$ ($f(\Sigma) < g(\Sigma)$) for all $\Sigma \in \mathcal{T}$ and $\mathcal{T}$ is specified only by the conditional independence relationships.*

The concept of almost qualitative comparison is similar and can be defined as follows:

Definition 6.2 *We say f is almost qualitatively larger (smaller) than g on $\mathcal{T} \subseteq \mathcal{S}^+$, if $f(\Sigma) > g(\Sigma)$ ($f(\Sigma) < g(\Sigma)$) for all $\Sigma \in \mathcal{T}$ and $\mathcal{T}$ is specified by both the conditional independence relationships and the sign of entries of Σ and some partial covariances.*

Notice that, the two definitions differ actually in the specification of the subset $\mathcal{T}$. Almost qualitative comparison is a qualitative comparison which is valid over a subset $\mathcal{T}$ of $\mathcal{S}^+$ specified by both conditional independence relationships and signs of the entries in the covariance matrix. Qualitative comparison, on the other hand, do not require the additional step of knowing the sign of the entries in covariance matrix and partial covariances.

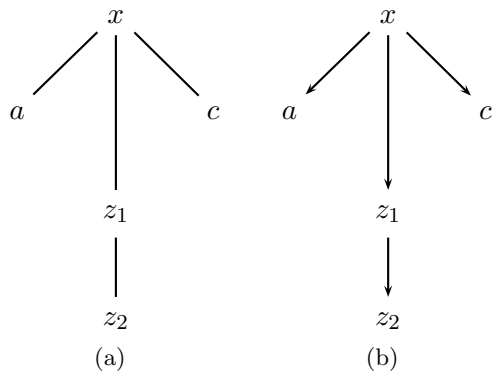


Figure 6.1 Graphical models satisfying the conditions of Theorem 6.1 and Corollary 6.1. In all cases $\rho^2_{ac} \geq \rho^2_{ac|z_2} \geq \rho^2_{ac|z_1}$.

The next proposition is trivial but will be heavily used in most of the proof below.

Proposition 6.1 *Let X, Y, Z be jointly Gaussian with Covariance Σ. Suppose $X \perp\!\!\!\perp Y \mid Z$. Then $\Sigma_{XY} = \Sigma_{XZ}\Sigma^{-1}_{ZZ}\Sigma_{ZY}$.* □

6.3 Some Key cases

We look at some key cases, which is discussed in details in Chaudhuri [2013]. These cases are not specific to any type of Graphical Models. We shall show later that more general cases can be reduced to these.

Here we hold two components a and c of V fixed. The variations in $\rho_{ac|Z}$, $\beta_{ac|Z}$ and $\beta_{ca|Z}$ are compared for different subsets Z of V. We do not do comparison with fixed conditionates as variations in partial correlations cannot be qualitatively compared. This is discussed in [Chaudhuri, 2013], Page 10.

Depending on the nature of pairwise unconditional association between a, c and the sets conditioned on, three situations may arise.

6.3.1 Situation 1

The components a, c, z_1 and z_2 are unconditionally pairwise dependent.

Theorem 6.1 *Suppose for some x, $a \perp\!\!\!\perp c \mid x$, $ac \perp\!\!\!\perp z_1 \mid x$ and $ac \perp\!\!\!\perp z_2 \mid z_1$. Then*

$$\sigma_{ac} \propto^{+} \sigma_{ac|z_1} \propto^{+} \sigma_{ac|z_2} \propto^{+} \sigma_{ax}\sigma_{xc}. \tag{6.3.1}$$

Theorem 6.1 shows that under the assumptions σ_{ac}, $\sigma_{ac|z_1}$ and $\sigma_{ac|z_2}$ all have the same sign determined by the signs of σ_{ax} and σ_{cx}. In view of Chaudhuri and Richardson [2003], Chaudhuri and Tan [2010], Chaudhuri [2013] the following result is immediate.

Corollary 6.1 *Under the conditions of Theorem 6.1, exactly one of the following holds*

(1) $\rho_{ac} \geq \rho_{ac|z_2} \geq \rho_{ac|z_1} \geq 0$, $\beta_{ac} \geq \beta_{ac|z_2} \geq \beta_{ac|z_1} \geq 0$ *and* $\beta_{ca} \geq \beta_{ca|z_2} \geq \beta_{ca|z_1} \geq 0$,

(2) $\rho_{ac} \leq \rho_{ac|z_2} \leq \rho_{ac|z_1} \leq 0$, $\beta_{ac} \leq \beta_{ac|z_2} \leq \beta_{ac|z_1} \leq 0$ *and* $\beta_{ca} \leq \beta_{ca|z_2} \leq \beta_{ca|z_1} \leq 0$.

Corollary 6.1 shows that the comparison between ρ_{ac}, $\rho_{ac|z_1}$, $\rho_{ac|z_2}$ is valid over $\mathcal{T}$ (see Definition 6.1) which is specified by the conditional independences in Theorem 6.1 and the sign of σ_{ac}. Thus by Definition 6.2 this comparison is almost qualitative not qualitative. Notice that, the two possibilities in Corollary 6.1 correspond to two disjoint subsets of $\mathcal{S}^{+}$.

The nature of the comparison depends on the sign of ρ_{ac} and not on the conditionates. Thus if a and c are observed, the comparison between these correlations can be easily made. The application to the comparison between the partial regression coefficients are similar.

The conditions of Theorem 6.1 may be satisfied by several Graphical models, including tree and polytree models. We present two examples in Figure 6.1 above. Note that, in this case x is allowed to be a, c or z_1.

6.3.2 Situation 2

The correlates a and c are independent, but both are dependent on the sets conditioned on.

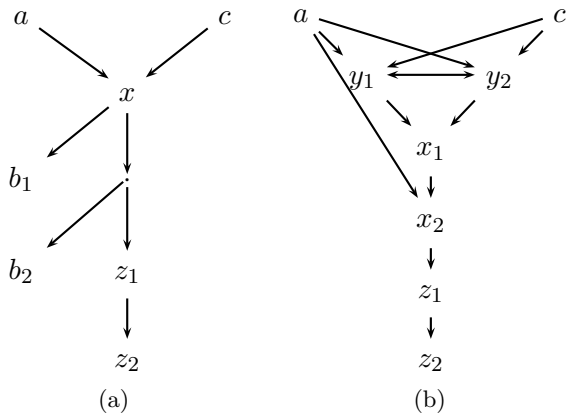


Figure 6.2 Graphical models satisfying the conditions of Theorem 6.2 and Corollary 6.2. In both cases $\rho^2_{ac|z_2} \leq \rho^2_{ac|z_1}$. Furthermore, in 6.2(a) $\rho^2_{ac|B} \leq \rho^2_{ac|Bz_2} \leq \rho^2_{ac|Bz_1}$ with $B = \{b_1, b_2\}$.

Theorem 6.2 *Suppose $a \perp\!\!\!\perp c$ and for some $x \in V \setminus \{a, c\}$, the condition $ac \perp\!\!\!\perp Bz_1 \mid x$ holds. Further, suppose that $z_2 \perp\!\!\!\perp acB \mid z_1$ holds. Then*

$$\sigma_{ac|x} \propto^+ \sigma_{ac|Bz_1} \propto^+ \sigma_{ac|Bz_2} \propto^+ -\sigma_{ax}\sigma_{xc}. \tag{6.3.2}$$

In this case the signs of $\sigma_{ac|x}$, $\sigma_{ac|Bz_1}$ and $\sigma_{ac|Bz_2}$ depend on the sign of the correlations between a and x and that between c and x. For direct acyclic graphs a collider x on the path between a and c would satisfy the conditions of Theorem 6.2. We shall show later (see Theorem 6.5) that the negative sign on the RHS of (6.3.2) depends on the number of colliders on the path.

In view of Chaudhuri and Richardson [2003], Chaudhuri and Tan [2010], Chaudhuri [2013] the following result can be derived.

Corollary 6.2 *Under the conditions of Theorem 6.2 one of the following holds.*

(1) $\rho_{ac|x} \geq \rho_{ac|Bz_1} \geq \rho_{ac|Bz_2} \geq 0$, $\beta_{ac|x} \geq \beta_{ac|Bz_1} \geq \beta_{ac|Bz_2} \geq 0$ *and* $\beta_{ca|x} \geq \beta_{ca|Bz_1} \geq \beta_{ca|Bz_2} \geq 0$,

(2) $\rho_{ac|x} \leq \rho_{ac|Bz_1} \leq \rho_{ac|Bz_2} \leq 0$, $\beta_{ac|x} \leq \beta_{ac|Bz_1} \leq \beta_{ac|Bz_2} \leq 0$, $\beta_{ca|x} \leq \beta_{ca|Bz_1} \leq \beta_{ca|Bz_2} \leq 0$.

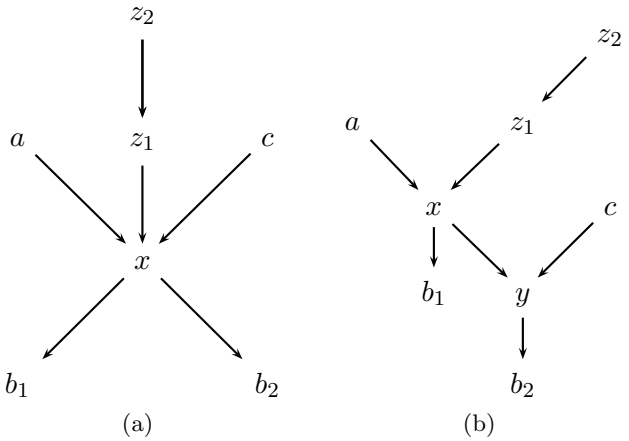


Figure 6.3 Graphical models satisfying the conditions of Theorem 6.3 and Corollary 6.3. In both cases $\rho^2_{ac|B} \leq \rho^2_{ac|Bz_2} \leq \rho^2_{ac|Bz_1}$ with $B = \{b_1, b_2\}$.

Note that, $\rho_{ac|x} \propto^+ \beta_{ac|x} \propto^+ \beta_{ca|x} \propto^+ -\sigma_{ax}\sigma_{xc}$. So from Corollary 6.2 it is clear that the nature of the comparison only depends on the sign of the product $\sigma_{ax}\sigma_{xc}$ and not on the conditionates B, z_1 ad z_2. We only need to observe a, c and x to determine the direction of inequalities in Corollary 6.2.

The vertex x is important on the path between a and c. Note that, unconditionally, a is independent of c. Thus, if we don't observe x, the observed covariance between a and c would be zero. However, a is not independent of c given x. This relation cannot be realized from the data unless x is observed.

In Theorem 6.2, x is specified by the conditional independence relationships and the sign of σ_{ax} and σ_{cx} in Σ. Alternatively, we can specify x with the sign of $\sigma_{ac|x}$. Thus there are two equivalent ways to represent x, where both require observation of x.

The conditions of Theorem 6.2 cannot be represented by an UG. However several other classes of models like directed acyclic graph and mixed ancestral graph can represent them. See Figure 6.2 for two examples.

6.3.3 Situation 3

At least one of a and c is independent of both the sets conditioned on.

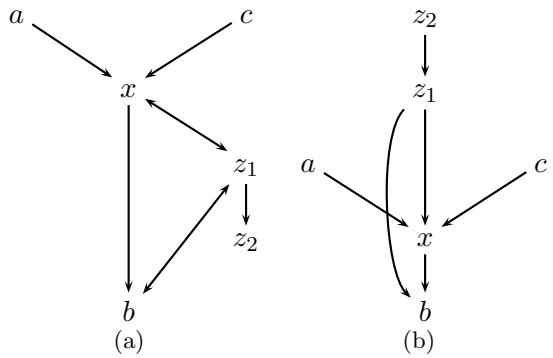


Figure 6.4 Graphical models satisfying the conditions of Theorem 6.3 and Corollary 6.3. In all cases $\rho^2_{ac|b} \leq \rho^2_{ac|bz_2} \leq \rho^2_{ac|bz_1}$.

Theorem 6.3 *Suppose $a \perp\!\!\!\perp z_1$. Let for some $x \in V \setminus \{a, z_1, z_2\}$, Σ satisfies one of the following two $((i), (ii))$ conditions:*

(i) $c \perp\!\!\!\perp az_1$, $z_2 \perp\!\!\!\perp acB \mid z_1$ and one of the following six conditions (a) $az_1 \perp\!\!\!\perp B \mid x$, (b) $az_1 \perp\!\!\!\perp B \mid cx$, (c) $cz_1 \perp\!\!\!\perp B \mid x$, (d) $cz_1 \perp\!\!\!\perp B \mid ax$, (e) $ac \perp\!\!\!\perp B \mid x$ and (f) $ac \perp\!\!\!\perp B \mid xz_1$ holds,

(ii) $az_1 \perp\!\!\!\perp cB \mid x$ and $z_2 \perp\!\!\!\perp acB \mid z_1$.

Then

$$\sigma_{ac|B} \propto^+ \sigma_{ac|Bz_1} \propto^+ \sigma_{ac|Bz_2}. \tag{6.3.3}$$

As before Theorem 6.3 leads to the following Corollary.

Corollary 6.3 *Under the conditions of Theorem 6.3 one of the following statements holds.*

(1) $\rho_{ac|Bz_1} \geq \rho_{ac|Bz_2} \geq \rho_{ac|B} \geq 0$, $\beta_{ac|Bz_1} \geq \beta_{ac|Bz_2} \geq \beta_{ac|B} \geq 0$ and $\beta_{ca|Bz_1} \geq \beta_{ca|Bz_2} \geq \beta_{ca|B} \geq 0$,

(2) $\rho_{ac|Bz_1} \leq \rho_{ac|Bz_2} \leq \rho_{ac|B} \leq 0$, $\beta_{ac|Bz_1} \leq \beta_{ac|Bz_2} \leq \beta_{ac|B} \leq 0$ and $\beta_{ca|Bz_1} \leq \beta_{ca|Bz_2} \leq \beta_{ca|B} \leq 0$.

The sign of $\sigma_{ac|B}$ determines the nature of comparison. Thus in this situation, unlike before the nature of comparison depends on the conditionate B. Theorem 6.3 does not

determine the sign of $\sigma_{ac|B}$. However, in many cases, eg. polytree models, Theorems 6.1 and 6.2 can be applied with each element of B as conditionates. From this the sign of $\sigma_{ac|B}$ can be determined, which will in turn determine the sign in Corollary 6.3.

The Graphical models satisfying the conditions of Theorem 6.3 can be varied. We present a few examples in Figures 6.3 and 6.4 above.

6.4 Applications to certain singly connected graphs

The results in Section 6.3 apply to any Gaussian random vector. In this section we discuss the implication of those results on a class of graphical Markov models, namely trees and polytrees. Our main motivation is to associate the rules of comparisons for the partial correlation and regression coefficients with the paths joining the correlates a and c and the paths connecting these correlates with the conditionates.

In order to make such association, we consider only graphs where any two vertices have at most one path joining them. These graphs are denoted by singly connected graphs. A tree and a forest are two obvious examples. Further, recall that for a directed graph G, its skeleton G^* can be obtained by replacing all directed edges by undirected ones. Our definition of path (See definition 3.5) does not take into account of its possible direction. Thus the directed graphs whose skeletons are trees are also singly connected. These graphs are denoted by polytrees.

For any singly connected graph, if we consider the correlates a and c and conditionate z , the paths ${}_a\pi_c$, ${}_a\pi_z$, ${}_c\pi_z$ has a unique intersection point $\mathfrak{n}(z)$. Chaudhuri [2005], Chaudhuri and Tan [2010] show that the nature of comparison between the squared partial correlations depend on the nature of three paths at $\mathfrak{n}(z)$.

Using them, we can compare all situations while dealing with tree and polytree. So the correlation with fixed conditionates can be almost qualitatively compared.

From the above definition, polytrees are clearly singly connected. This implies that there is one unique path between any two vertices, say a and c. We denote such a path as ${}_a\pi_c$. We start with a result which generalizes and has application to trees and polytrees with no colliders.

6.5 Applications to Gaussian Trees

For Gaussian tree models, for fixed a and c and for any conditionate z, $\mathfrak{n}(z)$ is a non-collider in all three paths ${}_a\pi_c$, ${}_a\pi_z$, ${}_c\pi_z$. Thus situation 1 holds and we can use Theorem 6.1 to almost qualitatively compare $\rho_{ac|z_1}$ and $\rho_{ac|z_2}$. Thus, under certain conditions imposed two sets of conditionates Z_1 and Z_2, we can qualitatively compare $\rho^2_{ac|Z_1}$ and $\rho^2_{ac|Z_2}$. It will be seen that the nature of comparison is determined by the sign of σ_{ac}. We first prove that the sign of $\sigma_{ac|Z}$ is the same as σ_{ac}.

Lemma 6.1 *Suppose for some x, z_1, z_2, ..., z_n, $a \perp\!\!\!\perp c \mid x$ and $ac \perp\!\!\!\perp z_1 z_2 \dots z_n \mid x$. Then*

$$\sigma_{ac|z_1 z_2 \dots z_n} \propto^+ \sigma_{ax}\sigma_{xc} \propto^+ \sigma_{ac}.$$

Proof: The proof is by induction on z_1, z_2, ..., z_n. For $k = 1$, from Theorem 6.1, clearly it follows that

$$\sigma_{ac|z_1} \propto^+ \sigma_{ax}\sigma_{xc} \propto^+ \sigma_{ac}.$$

Suppose the statement is true for $Z_k = \{z_1, z_2, \dots, z_k\}$, i.e.

$$\sigma_{ac|Z_k} \propto^+ \sigma_{ax}\sigma_{xc} \propto^+ \sigma_{ac}.$$

We show the result for $Z_{k+1} = Z_k \cup \{z_{k+1}\}$. Note that,

$$\sigma_{ac|Z_{k+1}} = \sigma_{ac|Z_k} - \frac{\sigma_{az_{k+1}|Z_k}\sigma_{cz_{k+1}|Z_k}}{\sigma_{z_{k+1}z_{k+1}|Z_k}}.$$

Notice that, since $a \perp\!\!\!\perp cZ_k|x$, we have $a \perp\!\!\!\perp c|xZ_k$. Now using Proposition 6.1 with $ac \perp\!\!\!\perp z_{k+1}|x$ and $a \perp\!\!\!\perp c|xZ_k$, we get

$$\sigma_{ac|Z_k} = \sigma_{ax|Z_k}\sigma_{cx|Z_k}/\sigma^2_{xx|Z_k}.$$
$$\sigma_{az_{k+1}|Z_k} = \sigma_{ax|Z_k}\sigma_{xz_{k+1}|Z_k}/\sigma_{xx|Z_k}.$$
$$\sigma_{cz_{k+1}|Z_k} = \sigma_{cx|Z_k}\sigma_{xz_{k+1}|Z_k}/\sigma_{xx|Z_k}.$$

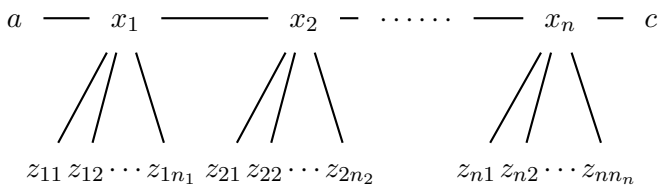


Figure 6.5 The tree discussed in Theorem 6.4.

By substitution, using Proposition 6.1 with $a \perp\!\!\!\perp c|Z_k x$, it follows that

$$\sigma_{ac|Z_{k+1}} = \frac{\sigma_{ax|Z_k}\sigma_{cx|Z_k}}{\sigma^2_{xx|Z_k}}\left[\sigma^2_{xx|Z_k} - \frac{\sigma^2_{xz_{k+1}|Z_k}}{\sigma_{z_{k+1}z_{k+1}|Z_k}}\right]$$
$$\propto^+ \sigma_{ax|Z_k}\sigma_{cx|Z_k}\sigma_{xx|Z_{k+1}} \propto^+ \sigma_{ac|Z_k}.$$

This completes the proof. □

A general Gaussian tree looks like Figure 6.5. Now, we apply Lemma 6.1 to general Gaussian Trees.

Theorem 6.4 *Suppose a and c are two vertices on Gaussian tree G. Let Z be the subset of vertices. Then*

$$\sigma_{ac|Z} \propto^+ \sigma_{ac}.$$

Proof: If there exist $z \in Z$ such that $z \in {}_a\pi_c$, then $a \perp\!\!\!\perp c|z$ and $\sigma_{ac|Z} = 0 \propto^+ \sigma_{ac}$. Therefore, it suffices to consider the case when $\forall z \in Z$, $z \notin {}_a\pi_c$. Now since G is a tree, for each $z \in Z$, there exist $\mathfrak{n}(z)$ such that $\mathfrak{n}(z) = {}_a\pi_c \cap {}_a\pi_z \cap {}_c\pi_z$. Suppose that there are n such vertex $\mathfrak{n}(z)$ on ${}_a\pi_c$. Further, let us enumerate these vertices as $x_1, ..., x_n$ as from their distance from a. For each x_i, let $z_i = \{z \in Z : \mathfrak{n}(z) = x_i\}$.

The proof is by induction. For $k = 1$, using Lemma 6.1 above, it is clear that $\sigma_{ac|z_1} \propto^+ \sigma_{ac}$.

Let $Z_k = \cup_{i=1}^{k} \cup_{j=1}^{n_i} \{z_{i,j}\}$ and $z_k = \{z_{k,1}, ..., z_{k,n_k}\}$. Suppose that the result hold for Z_k. That is, $\sigma_{ac|Z_k} \propto^+ \sigma_{ac}$. We show that $\sigma_{ac|Z_{k+1}} \propto^+ \sigma_{ac}$.

First of all note that, $\sigma_{ac|Z_{k+1}} = \sigma_{ac|Z_k} - \Sigma_{az_{k+1}|Z_k}\Sigma^{-1}_{z_{k+1}z_{k+1}|Z_k}\Sigma_{z_{k+1}c|Z_k}$.

From the graph, it is clear that $a \perp\!\!\!\perp c \mid x_{k+1}Z_k$. Using Proposition 6.1, it follows that

$$\sigma_{ac|Z_k} = \sigma_{x_{k+1}c|Z_k}\sigma_{ax_{k+1}|Z_k}/\sigma_{x_{k+1}x_{k+1}|Z_k}.$$

Further $a \perp\!\!\!\perp z_{k+1} \mid x_{k+1}Z_k$. Using Proposition 6.1, it follows that

$$\Sigma_{az_{k+1}|Z_k} = \Sigma_{x_{k+1}z_{k+1}|Z_k}\sigma_{ax_{k+1}|Z_k}/\sigma_{x_{k+1}x_{k+1}|Z_k}.$$

Moreover $c \perp\!\!\!\perp z_{k+1} \mid x_{k+1}Z_k$. This together with Proposition 6.1 gives

$$\Sigma_{z_{k+1}c|Z_k} = \sigma_{cx_{k+1}|Z_k}\Sigma_{z_{k+1}x_{k+1}|Z_k}/\sigma_{x_{k+1}x_{k+1}|Z_k}.$$

Now by substituting these relationship in the expression of $\sigma_{ac|Z_{k+1}}$ we get

$$\begin{aligned}
&\sigma_{ac|Z_{k+1}} \\
&= \frac{\sigma_{x_{k+1}c|Z_k}\sigma_{ax_{k+1}|Z_k}}{\sigma^2_{x_{k+1}x_{k+1}|Z_k}}\left(\sigma_{x_{k+1}x_{k+1}|Z_k} - \Sigma_{x_{k+1}z_{k+1}|Z_k}\Sigma^{-1}_{z_{k+1}z_{k+1}|Z_k}\Sigma_{z_{k+1}x_{k+1}|Z_k}\right) \\
&= \frac{\sigma_{x_{k+1}c|Z_k}\sigma_{ax_{k+1}|Z_k}}{\sigma^2_{x_{k+1}x_{k+1}|Z_k}}\sigma_{x_{k+1}x_{k+1}|z_{k+1}} \\
&\propto^+ \sigma_{ac|Z_k} \qquad (6.5.1)
\end{aligned}$$

where the last line holds using Proposition 6.1 with $a \perp\!\!\!\perp c|Z_k x_{k+1}$. By induction hypothesis, $\sigma_{ac|Z_k} \propto^+ \sigma_{ac}$. Therefore

$$\sigma_{ac|z_{k+1}} \propto^+ \sigma_{ac}.$$

This completes the proof. □

Theorem 6.4 shows that the sign of the conditional covariance of two correlates given the conditionates depend only on the sign of the unconditional covariance. For the special case of the tree in Figure 6.5, it is also straightforward to show that

$$\sigma_{ac} \propto^+ \sigma_{ax_1}\left(\prod_{i=1}^{n-1}\sigma_{x_i x_{i+1}}\right)\sigma_{x_n c}.$$

Theorem 6.4 can also be applied to polytree models with no collider, as they are Markov equivalent to tree models.

From Theorem 6.4 the following general result about Gaussian tree models can be derived. A significant step in the proof can be found in Chaudhuri [2013, Theorem 7]. We state the result for partial correlation. The result for partial regression coefficients are similar.

Corollary 6.4 *Suppose $G = (V, E)$ is a Gaussian tree, Suppose a and c are two vertices on G and let Z_1 and Z_2 are two subsets of V such that $ac \perp\!\!\!\perp Z_2 \mid Z_1$. Then either $0 \leq \rho_{ac|Z_1} \leq \rho_{ac|Z_2} \leq \rho_{ac}$ or $0 \geq \rho_{ac|Z_1} \geq \rho_{ac|Z_2} \geq \rho_{ac}$.*

Proof: Under the given assumptions, Chaudhuri [2013, Theorem 7] shows that

$$\rho^2_{ac|Z_1} \leq \rho^2_{ac|Z_2}.$$

Theorem 6.4 states that $\rho_{ac|Z_1} \propto^+ \rho_{ac}$ and $\rho_{ac|Z_2} \propto^+ \rho_{ac}$, Thus, Corollary 6.4 follows. □

6.6 Applications to Polytree Models

Now we turn to polytree models with colliders. A general polytree model can be found in Figure 6.9. We fix two correlates a and c on the graph and find the sign of $\sigma_{ac|Z}$. We show that this depends on the number and sign of covariances between the colliders on ${}_a\pi_c$ and also on the sign of the covariance between the correlates and their nearest colliders on ${}_a\pi_c$. Finally for two conditionates Z_1 and Z_2, we compare $\rho_{ac|Z_1}$ and $\rho_{ac|Z_2}$ almost qualitatively.

The qualitative comparison uses results from Chaudhuri [2005], who assumes that each vertex in the conditionate is d-connected to the path given the empty set. That is, for each $z \in Z$, the path ${}_a\pi_z \cap {}_c\pi_z$ does not have a collider. Even though this assumption is not crucial for determining the sign of $\sigma_{ac|Z}$, still we make this assumption. We also assume $z \cap {}_a\pi_c$ is empty for any conditionate z.

Since the graph is singly connected, for each $z \in Z$, there is unique vertex $\mathfrak{n}(z) =$

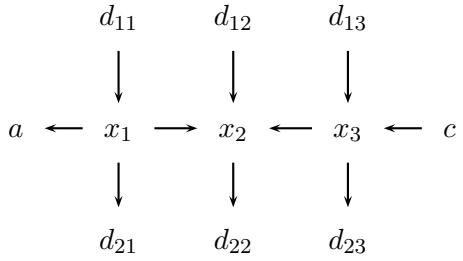


Figure 6.6 Example of a polytree. In this case, $\{d_{11}, d_{12}, d_{13}\} = \mathcal{D}^{(1)}_{ac}\}$, $\{d_{21}, d_{22}\} = \mathcal{D}^{(2)}_{ac}\}$ and $d_{31} = \mathcal{D}^{(3)}_{ac}$.

${}_a\pi_c \cap {}_a\pi_z \cap {}_c\pi_z$ on ${}_a\pi_c$. For two vertices x_i and x_j, we define

$$\mathcal{D}_{x_i x_j} = \left\{ z \in V : \exists x \in {}_{x_i}\pi_{x_j} \backslash \{x_i, x_j\} \text{ such that } x \in {}_{x_i}\pi_z \cap {}_{x_j}\pi_z \cap {}_{x_i}\pi_{x_j} \right\}.$$

Chaudhuri [2005] shows that depending on the nature of the paths ${}_a\pi_c$, ${}_a\pi_z$ and ${}_c\pi_z$ at $\mathfrak{n}(z)$, vertices in Z can be classified into three disjoint subsets. We describe these subsets below.

$$\begin{aligned}
\mathcal{D}^{(1)}_{x_i x_j} &= \{z \in \mathcal{D}_{x_i x_j} : \text{ at least one of the path } {}_{x_i}\pi_z, {}_{x_j}\pi_z \text{ has a collider at } \mathfrak{n}(z)\}. \\
\mathcal{D}^{(2)}_{x_i x_j} &= \{z \in \mathcal{D}_{x_i x_j} : {}_{x_i}\pi_z, {}_{x_j}\pi_z, {}_{x_i}\pi_{x_j} \text{ do not have collider at } \mathfrak{n}(z)\}. \\
\mathcal{D}^{(3)}_{x_i x_j} &= \{z \in \mathcal{D}_{x_i x_j} : \text{ Only } {}_{x_i}\pi_{x_j} \text{ has a collider at } \mathfrak{n}(z), {}_{x_i}\pi_z, {}_{x_j}\pi_z \\
&\qquad\qquad \text{do not have a collider at } \mathfrak{n}(z)\}.
\end{aligned}$$

An example of polytree is illustrated in Figure 6.6. Observe that for colliders such as x_2, $\{d_{22}\} = de(ch(x_2)) \in \mathcal{D}^{(3)}_{ac}$ and $\{d_{12}\} = an(pa(x_2)\backslash_a\pi_c) \in \mathcal{D}^{(1)}_{ac}$. For non-colliders such as x_1 and x_3, $\{d_{13}\} = an(pa(x_3)\backslash_a\pi_c) \in \mathcal{D}^{(1)}_{ac}$. For all other nodes, such as d_{21}, d_{22} and d_{23}, they are in $\mathcal{D}^{(2)}_{ac}$.

We now show a result on the sign of $\sigma_{ac|Z}$ on a polytree where there is no colliders on the path ${}_a\pi_c$.

Lemma 6.2 *Consider the graph in Figure 6.7. Let $a = x_0$ and $c = x_{n+1}$. Define $D^* = an(pa(a)\backslash_a\pi_c)$, $S_k = \mathcal{D}^{(1)}_{x_{k-1}x_{k+1}}$, $T_k = \mathcal{D}^{(2)}_{x_{k-1}x_{k+1}}$, $\mathcal{S} = \cup_{i=1}^n S_i$ and $\mathcal{T} = \cup_{i=1}^n T_i$.*

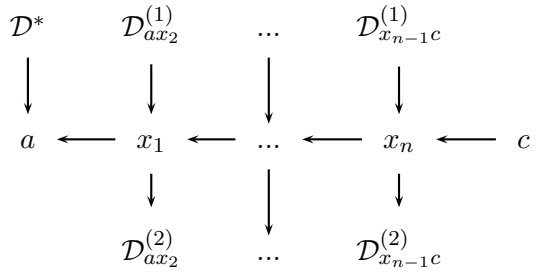


Figure 6.7 An example of a graph that satisfies the condition in Lemma 6.2. This graph structure can be found in Figure 6.8 between each "x_k and b_k" and "b_k and x_{k+1}".

We have

$$\sigma_{ac|\mathcal{S}\mathcal{T}D^*} \propto^+ \sigma_{ac}.$$

Proof: Since $D^* \perp\!\!\!\perp c|ST$, we have

$$\sigma_{ac|STD^*} = \sigma_{ac|ST} - \Sigma_{aD^*}\Sigma^{-1}_{D^*D^*|ST}\Sigma_{D^*c|ST} = \sigma_{ac|ST}$$

Therefore, it suffices to show that $\sigma_{ac|\mathcal{S}\mathcal{T}} \propto^+ \sigma_{ac}$. The proof is by induction.

For the case $\sigma_{ac|S_1T_1}$, since $S_1 \perp\!\!\!\perp c$ and using $a \perp\!\!\!\perp cS_1T_1|x_1$, $c \perp\!\!\!\perp T_1|x_1$ and $T_1 \perp\!\!\!\perp S_1|x_1$ with Proposition 6.1, we get

$$\begin{aligned}
\sigma_{ac|T_1} &= \sigma_{ac} - \Sigma_{aT_1}\Sigma^{-1}_{T_1T_1}\Sigma_{T_1c} && (6.6.1)\\
&= \frac{\sigma_{ax_1}\sigma_{x_1c}}{\sigma_{x_1x_1}} - \frac{\sigma_{ax_1}\Sigma_{x_1T_1}\Sigma^{-1}_{T_1T_1}\Sigma_{T_1x_1}\sigma_{x_1c}}{\sigma^2_{x_1x_1}}\\
&= \frac{\sigma_{ax_1}\sigma_{x_1c}}{\sigma_{x_1x_1}}\left[1 - \frac{\Sigma_{x_1T_1}\Sigma^{-1}_{T_1T_1}\Sigma_{T_1x_1}}{\sigma_{x_1x_1}}\right]\\
\Sigma_{aS_1|T_1} &= \Sigma_{aS_1} - \Sigma_{aT_1}\Sigma^{-1}_{T_1T_1}\Sigma_{T_1S_1} && (6.6.2)\\
&= \frac{\sigma_{ax_1}\Sigma_{x_1S_1}}{\sigma_{x_1x_1}} - \frac{\Sigma_{ax_1}\Sigma_{x_1T_1}\Sigma^{-1}_{T_1T_1}\Sigma_{T_1x_1}\Sigma_{x_1S_1}}{\sigma^2_{x_1x_1}}\\
&= \frac{\sigma_{ax_1}}{\sigma_{x_1x_1}}\left[1 - \frac{\Sigma_{x_1T_1}\Sigma^{-1}_{T_1T_1}\Sigma_{T_1x_1}}{\sigma_{x_1x_1}}\right]\Sigma_{x_1S_1}\\
\Sigma_{S_1c|T_1} &= 0 - \Sigma_{S_1T_1}\Sigma^{-1}_{T_1T_1}\Sigma_{T_1c} && (6.6.3)
\end{aligned}$$

$$= -\frac{\Sigma_{S_1x_1}\Sigma_{x_1T_1}\Sigma^{-1}_{T_1T_1}\Sigma_{T_1x_1}\sigma_{x_1c}}{\sigma^2_{x_1x_1}}.$$

Using equation (6.6.1), (6.6.2) and (6.6.3), since $\Sigma^{-1}_{S_1S_1|T_1}$ and $\Sigma^{-1}_{T_1T_1}$ is positive definite and using $a \perp\!\!\!\perp c|x_1$ with Proposition 6.1, we get

$$\begin{aligned}
\sigma_{ac|S_1T_1} &= \sigma_{ac|T_1} - \Sigma_{aS_1|T_1}\Sigma^{-1}_{S_1S_1|T_1}\Sigma_{S_1c|T_1} \\
&= \frac{\sigma_{ax_1}\sigma_{x_1c}}{\sigma_{x_1x_1}}\left[1 - \frac{\Sigma_{x_1T_1}\Sigma^{-1}_{T_1T_1}\Sigma_{T_1x_1}}{\sigma_{x_1x_1}}\right]\left[1 + \frac{\Sigma_{x_1S}\Sigma^{-1}_{S_1S_1|T_1}\Sigma_{S_1x_1}\Sigma_{x_1T_1}\Sigma^{-1}_{T_1T_1}\Sigma_{T_1x_1}}{\sigma^2_{x_1x_1}}\right] \\
&\propto^+ \frac{\sigma_{ax_1}\sigma_{x_1c}}{\sigma_{x_1x_1}}\left[1 - \frac{\Sigma_{x_1T_1}\Sigma^{-1}_{T_1T_1}\Sigma_{T_1x_1}}{\sigma_{x_1x_1}}\right] \\
&= \frac{\sigma_{ax_1}\sigma_{x_1c}}{\sigma^2_{x_1x_1}}\left[\sigma_{x_1x_1} - \Sigma_{x_1T_1}\Sigma^{-1}_{T_1T_1}\Sigma_{T_1x_1}\right] \\
&\propto^+ \sigma_{ac}\sigma_{x_1x_1|T_1} \propto^+ \sigma_{ac}.
\end{aligned}$$

Now for any integer k, let $\mathcal{S}_k = \cup_{i=1}^k S_i$, $\mathcal{T}_k = \cup_{i=1}^k T_i$ and $U_k = \mathcal{S}_k \cup \mathcal{T}_k$. Let $\mathcal{Q}_{ac} = \Sigma_{aT_{k+1}|U_k}\Sigma^{-1}_{T_{k+1}T_{k+1}|U_k}\Sigma_{T_{k+1}c|U_k}$, $\mathcal{Q}_{x_{k+1}x_{k+1}} = \Sigma_{x_{k+1}T_{k+1}|U_k}\Sigma^{-1}_{T_{k+1}T_{k+1}|U_k}\Sigma_{T_{k+1}x_{k+1}|U_k}$. Assume that $\sigma_{ac|U_k} \propto^+ \sigma_{ac}$, we show that for $U_{k+1} = \mathcal{S}_k \cup S_{k+1} \cup \mathcal{T}_k \cup T_{k+1}$, $\sigma_{ac|U_{k+1}} \propto^+ \sigma_{ac}$. Now, since $a \perp\!\!\!\perp cS_{k+1}T_{k+1}|x_{k+1}U_k$, $c \perp\!\!\!\perp T_{k+1}|x_{k+1}U_k$ and $T_{k+1} \perp\!\!\!\perp S_{k+1}|x_{k+1}U_k$, together with Proposition 6.1, we get

$$\begin{aligned}
\sigma_{ac|\mathcal{S}_k\mathcal{T}_{k+1}} &= \sigma_{ac|U_k} - \mathcal{Q}_{ac} \qquad (6.6.4)\\
&= \frac{\sigma_{ax_{k+1}|U_k}\sigma_{x_{k+1}c|U_k}}{\sigma_{x_{k+1}x_{k+1}|U_k}} - \frac{\sigma_{ax_{k+1}|U_k}\mathcal{Q}_{x_{k+1}x_{k+1}}\sigma_{x_{k+1}c|U_k}}{\sigma^2_{x_{k+1}x_{k+1}|U_k}} \\
&= \frac{\sigma_{ax_{k+1}|U_k}\sigma_{x_{k+1}c|U_k}}{\sigma_{x_{k+1}x_{k+1}|U_k}}\left[1 - \frac{\mathcal{Q}_{x_{k+1}x_{k+1}}}{\sigma_{x_{k+1}x_{k+1}|U_k}}\right] \\
\Sigma_{aS_{k+1}|\mathcal{S}_k\mathcal{T}_{k+1}} &= \Sigma_{aS_{k+1}|U_k} - \Sigma_{aT_{k+1}|U_k}\Sigma^{-1}_{T_{k+1}T_{k+1}|U_k}\Sigma_{T_{k+1}S_{k+1}|U_k} \qquad (6.6.5)\\
&= \frac{\sigma_{ax_{k+1}|U_k}\Sigma_{x_{k+1}S_{k+1}|U_k}}{\sigma_{x_{k+1}x_{k+1}|U_k}} - \frac{\sigma_{ax_{k+1}|U_k}\mathcal{Q}_{x_{k+1}x_{k+1}}\Sigma_{x_{k+1}S_{k+1}|U_k}}{\sigma^2_{x_{k+1}x_{k+1}|U_k}} \\
&= \frac{\sigma_{ax_{k+1}|U_k}}{\sigma_{x_{k+1}x_{k+1}|U_k}}\left[1 - \frac{\mathcal{Q}_{x_{k+1}x_{k+1}}}{\sigma_{x_{k+1}x_{k+1}|U_k}}\right]\Sigma_{x_{k+1}S_{k+1}|U_k} \\
\Sigma_{S_{k+1}c|\mathcal{S}_k\mathcal{T}_{k+1}} &= 0 - \Sigma_{S_{k+1}T_{k+1}|U_k}\Sigma^{-1}_{T_{k+1}T_{k+1}|U_k}\Sigma_{T_{k+1}c|U_k} \qquad (6.6.6)
\end{aligned}$$

$$= -\frac{\Sigma_{S_{k+1}x_{k+1}|U_k}\mathcal{Q}_{x_{k+1}x_{k+1}}\sigma_{x_{k+1}c|U_k}}{\sigma^2_{x_{k+1}x_{k+1}|U_k}}.$$

Therefore, substitution of (6.6.4), (6.6.5) and (6.6.6) with the fact that $\Sigma^{-1}_{S_{k+1}S_{k+1}|\mathcal{S}_kT}$ and $\Sigma^{-1}_{T_{k+1}T_{k+1}|U_k}$ is positive definite, we get

$$\begin{aligned}
\sigma_{ac|\mathcal{S}_{k+1}\mathcal{T}_{k+1}} &= \sigma_{ac|\mathcal{S}_k\mathcal{T}_{k+1}} - \Sigma_{aS_{k+1}|\mathcal{S}_k\mathcal{T}_{k+1}}\Sigma^{-1}_{S_{k+1}S_{k+1}|\mathcal{S}_k\mathcal{T}_{k+1}}\Sigma_{S_{k+1}c|\mathcal{S}_k\mathcal{T}_{k+1}} \\
&= \frac{\sigma_{ax_{k+1}|U_k}\sigma_{x_{k+1}c|U_k}}{\sigma_{x_{k+1}x_{k+1}|U_k}}\left[1 - \frac{\mathcal{Q}_{x_{k+1}x_{k+1}}}{\sigma_{x_{k+1}x_{k+1}|U_k}}\right] + \frac{\sigma_{ax_{k+1}|U_k}}{\sigma_{x_{k+1}x_{k+1}|U_k}}\left[1 - \frac{\mathcal{Q}_{x_{k+1}x_{k+1}}}{\sigma_{x_{k+1}x_{k+1}|U_k}}\right] \\
&\qquad \Sigma_{x_{k+1}S_{k+1}|U_k}\Sigma^{-1}_{S_{k+1}S_{k+1}|\mathcal{S}_k\mathcal{T}_{k+1}}\frac{\Sigma_{S_{k+1}x_{k+1}|U_k}\mathcal{Q}_{x_{k+1}x_{k+1}}\sigma_{x_{k+1}c|U_k}}{\sigma^2_{x_{k+1}x_{k+1}|U_k}} \\
&= \sigma_{ac|\mathcal{S}_k\mathcal{T}_{k+1}}\left[1 + \frac{\Sigma_{x_{k+1}S_{k+1}|U_k}\Sigma^{-1}_{S_{k+1}S_{k+1}|\mathcal{S}_kT}\Sigma_{S_{k+1}x_{k+1}|U_k}\mathcal{Q}_{x_{k+1}x_{k+1}}}{\sigma^2_{x_{k+1}x_{k+1}|U_k}}\right] \\
&\propto^+ \sigma_{ac|\mathcal{S}_k\mathcal{T}_{k+1}}.
\end{aligned}$$

Also, using the induction hypothesis, (6.6.4) and $a \perp\!\!\!\perp c|x_{k+1}U_k$ with Proposition 6.1, we get

$$\begin{aligned}
\sigma_{ac|\mathcal{S}_k\mathcal{T}_{k+1}} &= \frac{\sigma_{ax_{k+1}|U_k}\sigma_{x_{k+1}c|U_k}}{\sigma^2_{x_{k+1}x_{k+1}|U_k}}\left[\sigma_{x_{k+1}x_{k+1}|U_k} - \mathcal{Q}_{x_{k+1}x_{k+1}}\right] \\
&= \sigma_{ac|U_k}\sigma_{x_{k+1}x_{k+1}|\mathcal{S}_kT} \\
&\propto^+ \sigma_{ac|U_k} \propto^+ \sigma_{ac}.
\end{aligned}$$

□

Note that Lemma 6.2 generalises Theorem 6.3 (ii) for polytree models.

We now show a very general result on the sign of $\sigma_{ac|Z}$. For any two vertices a and c on a polytree, in most general situation ${}_a\pi_c$ would have several (say n) colliders $\{x_1, ..., x_n\}$. If the conditionate $Z \cap ch(x_i) = \phi$, for any $i = 1, ..., n$, $\sigma_{ac|Z} = 0$. We now show that for any conditionate z such that for each $z \in Z$ is d-connected to some node

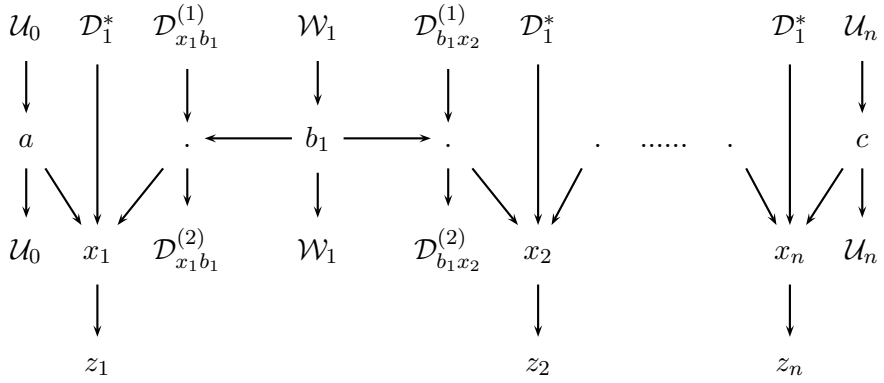


Figure 6.8 The polytree discussed in Theorem 6.5.

on ${}_a\pi_c$ given the empty set,

$$\sigma_{ac|Z} \propto^+ (-1)^n \sigma_{ax_1}\sigma_{x_nc}\prod_{k=1}^{n-1}\sigma_{x_kx_{k+1}}. \tag{6.6.7}$$

Note that n is the number of colliders on ${}_a\pi_c$. The result in (6.6.7) is developed over the next two theorems. In the first, we assume that each collider x_i on ${}_a\pi_c$ has exactly one descendant in the conditionate Z. Then we show that if there are more than one descendant of a collider in Z, we can condition on any one of them and the sign won't change.

Theorem 6.5 *Consider the DAG in Figure 6.8. Set* $a = x_0$, $c = x_{n+1}$. *Let* $\mathcal{W}_k = \mathcal{D}^{(2)}_{x_kx_{k+1}} \setminus \left(\mathcal{D}^{(2)}_{x_kb_k} \cup \mathcal{D}^{(2)}_{b_1x_2}\right)$, $D_k^* = an(x_k)\setminus\left(\mathcal{D}_{x_{k-1}x_k} \cup \mathcal{D}_{x_kx_{k+1}} \cup {}_a\pi_c\right)$, $\mathcal{Z}_k = ch(x_k)$ *and* $|\mathcal{Z}_k| = 1$. *Furthermore, let* $\mathcal{U}_k = \left(\mathcal{D}^{(1)}_{x_kx_{k+1}} \cup \mathcal{D}^{(2)}_{x_kx_{k+1}} \cup \mathcal{D}^*_k \cup \mathcal{D}^*_{k+1}\right)\setminus\mathcal{W}_k$ *for* $k = 1, ..., n-1$, $\mathcal{U}_0 = an\left(pa(a)\right) \cup de\left(ch(a)\setminus_a\pi_c\right)$ *and* $\mathcal{U}_n = an\left(pa(c)\right) \cup de\left(ch(c)\setminus_a\pi_c\right)$. *Then*

$$\sigma_{ac|\mathcal{U}\mathcal{W}\mathcal{Z}} \propto^+ (-1)^n \sigma_{ax_1}\sigma_{x_nc}\prod_{k=1}^{n-1}\sigma_{x_kx_{k+1}}$$

where $\mathcal{U} = \cup_{i=0}^{n}\mathcal{U}_k$, $\mathcal{W} = \cup_{i=1}^{n-1}\mathcal{W}_k$ *and* $\mathcal{Z} = \cup_{i=1}^{n-1}\mathcal{Z}_k$.

Proof: The proof consists of three parts.

(1) First, we show that

$$\sigma_{ac|\mathcal{U}\mathcal{W}\mathcal{Z}} \propto^{+} (-1)^n \sigma_{ax_1}\sigma_{x_nc}\prod_{k=1}^{n-1}\sigma_{x_kx_{k+1}|\mathcal{U}_k\mathcal{W}_k}.$$

Using Proposition 6.1 with $z_k \perp\!\!\!\perp z_{k+1}|x_k$, we get

$$\sigma_{z_kz_{k+1}|\mathcal{W}\mathcal{U}} = \frac{\sigma_{z_kx_k|\mathcal{W}\mathcal{U}}\sigma_{x_kz_{k+1}|\mathcal{W}\mathcal{U}}}{\sigma_{x_kx_k||\mathcal{W}\mathcal{U}}}.$$

Therefore, it is clear that

$$\sigma_{z_kz_{k+1}|\mathcal{W}\mathcal{U}} \propto^{+} \sigma_{z_kx_k|\mathcal{W}\mathcal{U}}\sigma_{x_kz_{k+1}|\mathcal{W}\mathcal{U}}. \tag{6.6.8}$$

Similarly, applying $x_k \perp\!\!\!\perp z_{k+1}|x_{k+1}$ and Proposition 6.1 on $\sigma_{x_kz_{k+1}|\mathcal{W}\mathcal{U}}$ in (6.6.8) gives us

$$\sigma_{z_kz_{k+1}|\mathcal{W}\mathcal{U}} \propto^{+} \sigma_{z_kx_k|\mathcal{W}\mathcal{U}}\sigma_{x_kx_{k+1}|\mathcal{W}\mathcal{U}}\sigma_{x_{k+1}z_{k+1}|\mathcal{W}\mathcal{U}} \tag{6.6.9}$$

Since $\sigma_{z_kz_j|\mathcal{W}\mathcal{U}} = 0$ for $j \neq \{k, k+1, k-1\}$, it is clear that $\Sigma_{\mathcal{Z}\mathcal{Z}|\mathcal{W}\mathcal{U}}$ is a tridiagonal matrix. The $(1,n)$ entry of $\Sigma^{-1}_{\mathcal{Z}\mathcal{Z}|\mathcal{W}\mathcal{U}}$ is given by

$$(-1)^{n+1}det(\Sigma_{\mathcal{Z}\mathcal{Z}|\mathcal{W}\mathcal{U}})\begin{vmatrix}\sigma_{z_2z_1|\mathcal{W}\mathcal{U}} & & & \\ 0 & \ddots & & \\ \vdots & \ddots & \ddots & \\ 0 & \cdots & 0 & \sigma_{z_{n-1}z_n|\mathcal{W}\mathcal{U}}\end{vmatrix}$$

where $det(\Sigma_{\mathcal{Z}\mathcal{Z}|\mathcal{W}\mathcal{U}})$ denotes the determinant of $\Sigma_{\mathcal{Z}\mathcal{Z}|\mathcal{W}\mathcal{U}}$. Since $\Sigma_{\mathcal{Z}\mathcal{Z}|\mathcal{W}\mathcal{U}}$ is positive definite, its determinant is positive. Therefore, since $a \perp\!\!\!\perp c|\mathcal{W}\mathcal{U}$, using (6.6.9), we

get

$$\begin{aligned}
\sigma_{ac|\mathcal{U}\mathcal{W}\mathcal{Z}} &= \sigma_{ac|\mathcal{W}\mathcal{U}} - \Sigma_{a\mathcal{Z}|\mathcal{W}\mathcal{U}}\Sigma^{-1}_{\mathcal{Z}\mathcal{Z}|\mathcal{W}\mathcal{U}}\Sigma_{\mathcal{Z}c|\mathcal{W}\mathcal{U}} \\
&= -\begin{pmatrix} \sigma_{az_1|\mathcal{W}\mathcal{U}} \\ 0 \\ \vdots \\ 0 \end{pmatrix}^T \Sigma^{-1}_{\mathcal{Z}\mathcal{Z}|\mathcal{W}\mathcal{U}} \begin{pmatrix} 0 \\ \vdots \\ 0 \\ \sigma_{cz_n|\mathcal{W}\mathcal{U}} \end{pmatrix} \\
&\propto^+ (-1)^{n+2}\sigma_{az_1|\mathcal{W}\mathcal{U}}\sigma_{z_1z_2|\mathcal{W}\mathcal{U}}\dots\sigma_{z_{n-1}z_n|\mathcal{W}\mathcal{U}}\sigma_{z_nc|\mathcal{W}\mathcal{U}} \\
&\propto^+ (-1)^{n}\sigma_{ax_1|\mathcal{W}\mathcal{U}}\sigma^2_{x_1z_1|\mathcal{W}\mathcal{U}}\sigma_{x_1x_2|\mathcal{W}\mathcal{U}}\sigma^2_{x_2z_2|\mathcal{W}\mathcal{U}}\dots\sigma_{x_{n-1}x_n|\mathcal{W}\mathcal{U}}\sigma^2_{x_nz_n|\mathcal{W}\mathcal{U}}\sigma_{x_nc|\mathcal{U}\mathcal{W}} \\
&\propto^+ (-1)^{n}\sigma_{ax_1|\mathcal{W}\mathcal{U}}\sigma_{x_1x_2|\mathcal{W}\mathcal{U}}\dots\sigma_{x_{n-1}x_n|\mathcal{W}\mathcal{U}}\sigma_{x_nc|\mathcal{W}\mathcal{U}}. \qquad (6.6.10)
\end{aligned}$$

Now, since $a \perp\!\!\!\perp \mathcal{W}_1, ..., \mathcal{W}_{n-1}, \mathcal{U}_1, ..., \mathcal{U}_n$ and using $x_1 \perp\!\!\!\perp \mathcal{U}_0|a$ with Proposition 6.1, we get

$$\begin{aligned}
\sigma_{ax_1|\mathcal{W}\mathcal{U}} = \sigma_{ax_1|\mathcal{U}_0} &= \sigma_{ax_1} - \Sigma_{a\mathcal{U}_0}\Sigma^{-1}_{\mathcal{U}_0\mathcal{U}_0}\Sigma_{\mathcal{U}_0x_1} \\
&= \sigma_{ax_1} - \Sigma_{a\mathcal{U}_0}\Sigma^{-1}_{\mathcal{U}_0\mathcal{U}_0}\Sigma_{\mathcal{U}_0a}\sigma^{-1}_{aa}\sigma_{ax_1} \\
&= \frac{\sigma_{ax_1}}{\sigma^2_{aa}}\left[\sigma_{aa} - \Sigma_{a\mathcal{U}_0}\Sigma^{-1}_{\mathcal{U}_0\mathcal{U}_0}\Sigma_{\mathcal{U}_0a}\right] \propto^+ \sigma_{ax_1}.
\end{aligned}$$

Similar, since $c \perp\!\!\!\perp \mathcal{W}_1, ..., \mathcal{W}_{n-1}, \mathcal{U}_0, ..., \mathcal{U}_{n-1}$ and using $x_n \perp\!\!\!\perp \mathcal{U}_n|c$ with Proposition 6.1, we get

$$\begin{aligned}
\sigma_{x_nc|\mathcal{W}\mathcal{U}} = \sigma_{cx_n|\mathcal{U}_n} &= \sigma_{cx_n} - \Sigma_{c\mathcal{U}_n}\Sigma^{-1}_{\mathcal{U}_n\mathcal{U}_n}\Sigma_{\mathcal{U}_nx_n} \\
&= \sigma_{cx_n} - \Sigma_{c\mathcal{U}_n}\Sigma^{-1}_{\mathcal{U}_n\mathcal{U}_n}\Sigma_{\mathcal{U}_nc}\sigma^{-1}_{cc}\sigma_{cx_n} \\
&= \frac{\sigma_{cx_n}}{\sigma^2_{cc}}\left[\sigma_{cc} - \Sigma_{c\mathcal{U}_n}\Sigma^{-1}_{\mathcal{U}_n\mathcal{U}_n}\Sigma_{\mathcal{U}_nc}\right] \propto^+ \sigma_{cx_n}.
\end{aligned}$$

Also, note that $x_{k+1} \perp\!\!\!\perp \mathcal{W}_1 \dots \mathcal{W}_{k-1}, \mathcal{U}_1 \dots \mathcal{U}_{k-1}$ and $x_k \perp\!\!\!\perp \mathcal{W}_{k+1} \dots \mathcal{W}_n, \mathcal{U}_{k+1} \dots \mathcal{U}_n$. Therefore, we get

$$\sigma_{x_kx_{k+1}|\mathcal{W}\mathcal{U}} = \sigma_{x_kx_{k+1}|\mathcal{W}_k\mathcal{U}_k}.$$

This completes the first part.

(2) Note that for consecutive colliders x_k, x_{k+1}, there is a vertex $b_k = {}_a\pi_c \cap an(x_k) \cap an(x_{k+1})$. Clearly b_k is a non-collider on the path ${}_a\pi_c$. Next we show that

$$\sigma_{x_k x_{k+1}|\mathcal{U}\mathcal{W}} \propto^+ \sigma_{x_k b_k} \sigma_{b_k x_{k+1}}.$$

Now, using $x_k \perp\!\!\!\perp x_{k+1} \mid \mathcal{U}_k b_k, x_k \perp\!\!\!\perp \mathcal{W}_k | b_k \mathcal{U}_k$, $\mathcal{W}_k \perp\!\!\!\perp x_{k+1} | \mathcal{U}_k b_k$ and Proposition 6.1, we get $\sigma_{x_k x_{k+1}|\mathcal{U}_k} = \frac{\sigma_{x_k b_k|\mathcal{U}_k} \sigma_{b_k x_{k+1}|\mathcal{U}_k}}{\sigma_{b_k b_k|\mathcal{U}_k}}$, $\Sigma_{x_k \mathcal{W}_k|\mathcal{U}_k} = \frac{\sigma_{x_k b_k|\mathcal{U}_k} \Sigma_{b_k \mathcal{W}_k|\mathcal{U}_k}}{\sigma_{b_k b_k|\mathcal{U}_k}}$ and $\Sigma_{\mathcal{W}_k x_{k+1}|\mathcal{U}_k} = \frac{\Sigma_{\mathcal{W}_k b_k|\mathcal{U}_k} \sigma_{b_k x_{k+1}|\mathcal{U}_k}}{\sigma_{b_k b_k|\mathcal{U}_k}}$. Using these relations, we get

$$\begin{aligned}
\sigma_{x_k x_{k+1}|\mathcal{U}_k \mathcal{W}_k} &= \sigma_{x_k x_{k+1}|\mathcal{U}_k} - \Sigma_{x_k \mathcal{W}_k|\mathcal{U}_k} \Sigma^{-1}_{\mathcal{W}_k \mathcal{W}_k|\mathcal{U}_k} \Sigma_{\mathcal{W}_k x_{k+1}|\mathcal{U}_k} \\
&= \frac{\sigma_{x_k b_k|\mathcal{U}_k} \sigma_{b_k x_{k+1}|\mathcal{U}_k}}{\sigma_{b_k b_k|\mathcal{U}_k}} - \frac{\sigma_{x_k b_k|\mathcal{U}_k} \Sigma_{b_k \mathcal{W}_k|\mathcal{U}_k} \Sigma^{-1}_{\mathcal{W}_k \mathcal{W}_k|\mathcal{U}_k} \Sigma_{\mathcal{W}_k b_k|\mathcal{U}_k} \sigma_{b_k x_{k+1}|\mathcal{U}_k}}{\sigma^2_{b_k b_k|\mathcal{U}_k}} \\
&= \frac{\sigma_{x_k b_k|\mathcal{U}_k} \sigma_{b_k x_{k+1}|\mathcal{U}_k}}{\sigma^2_{b_k b_k|\mathcal{U}_k}} (\sigma_{b_k b_k|\mathcal{U}_k} - \Sigma_{b_k \mathcal{W}_k|\mathcal{U}_k} \Sigma^{-1}_{\mathcal{W}_k \mathcal{W}_k|\mathcal{U}_k} \Sigma_{\mathcal{W}_k b_k|\mathcal{U}_k}) \\
&= \frac{\sigma_{x_k b_k|\mathcal{U}_k} \sigma_{b_k x_{k+1}|\mathcal{U}_k}}{\sigma^2_{b_k b_k|\mathcal{U}_k}} (\sigma_{b_k b_k|\mathcal{U}_k \mathcal{W}_k}) \\
&\propto^+ \sigma_{x_k b_k|\mathcal{U}_k} \sigma_{b_k x_{k+1}|\mathcal{U}_k}.
\end{aligned}$$

Now, notice that the structures from x_k to b_k and x_{k+1} to b_k are similar. Therefore, it suffices to show that $\sigma_{x_k b_k|\mathcal{U}_k} \propto^+ \sigma_{x_k b_k}$.

Let $\mathcal{V}_1 = \mathcal{D}^{(2)}_{b_k x_{k+1}}$, $\mathcal{V}_2 = \mathcal{U}_k \backslash \mathcal{V}_1$ and $\mathcal{V}_3 = \mathcal{V}_2 \backslash \left(\mathcal{D}^{(1)}_{b_k x_{k+1}} \cup D^*_{k+1} \right)$. Using Lemma 6.2, we have

$$\sigma_{x_k b_k|\mathcal{V}_3} \propto^+ \sigma_{x_k b_k}. \tag{6.6.11}$$

Using $x_k \perp\!\!\!\perp \mathcal{V}_1 | \mathcal{V}_2 b_k$ with Proposition 6.1, we get

$$\Sigma_{x_k \mathcal{V}_1|\mathcal{V}_2} = \frac{\sigma_{x_k b_k|\mathcal{V}_2} \Sigma_{b_k \mathcal{V}_1|\mathcal{V}_2}}{\sigma_{b_k b_k|\mathcal{V}_2}} \tag{6.6.12}$$

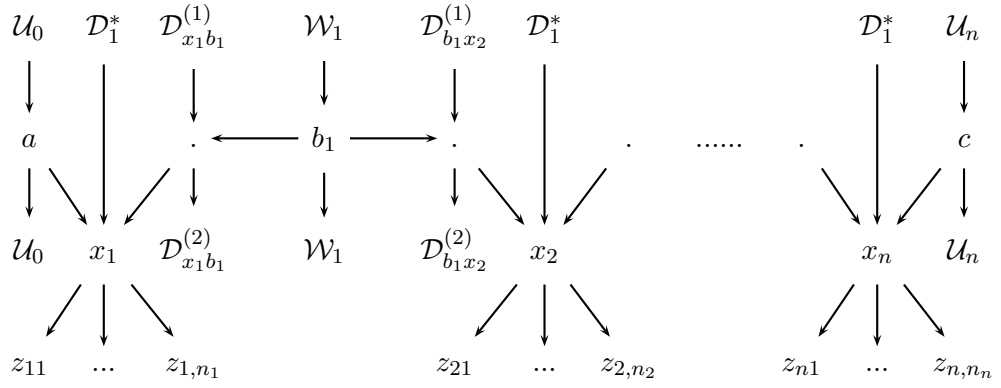


Figure 6.9 A polytree with multiple descendents on each x_k

Using that (6.6.11), (6.6.12) and noting that $b_k \perp\!\!\!\perp \mathcal{D}_{b_k x_{k+1}}^{(1)} \cup D_{k+1}^* | \mathcal{V}_3$, , we get

$$\begin{aligned}
\sigma_{x_k b_k|\mathcal{U}_k} &= \sigma_{x_k b_k|\mathcal{V}_2} - \Sigma_{x_k \mathcal{V}_1|\mathcal{V}_2} \Sigma^{-1}_{\mathcal{V}_1\mathcal{V}_1|\mathcal{V}_2} \Sigma_{\mathcal{V}_1 b_k|\mathcal{V}_2} \\
&= \sigma_{x_k b_k|\mathcal{V}_2} - \frac{\sigma_{x_k b_k|\mathcal{V}_2} \Sigma_{b_k \mathcal{V}_1|\mathcal{V}_2}}{\sigma_{b_k b_k|\mathcal{V}_2}} \Sigma^{-1}_{\mathcal{V}_1\mathcal{V}_1|\mathcal{V}_2} \Sigma_{\mathcal{V}_1 b_k|\mathcal{V}_2} \\
&= \frac{\sigma_{x_k b_k|\mathcal{V}_2}}{\sigma_{b_k b_k|\mathcal{V}_2}} (\sigma_{b_k b_k|\mathcal{V}_2} - \Sigma_{b_k \mathcal{V}_1|\mathcal{V}_2} \Sigma^{-1}_{\mathcal{V}_1\mathcal{V}_1|\mathcal{V}_2} \Sigma_{\mathcal{V}_1 b_k|\mathcal{V}_2}) \\
&\propto^+ \sigma_{x_k b_k|\mathcal{V}_2} \\
&= \sigma_{x_k b_k|\mathcal{V}_3} \propto^+ \sigma_{x_k b_k}.
\end{aligned}$$

(3) Finally, we want to show that $\sigma_{x_k b_k} \sigma_{b_k x_{k+1}} \propto^+ \sigma_{x_k x_{k+1}}$. This follows from using $x_k \perp\!\!\!\perp x_{k+1}|b_k$ with Proposition 6.1, which states that

$$\sigma_{x_k x_{k+1}} = \frac{\sigma_{x_k b_k} \sigma_{b_k x_{k+1}}}{\sigma_{b_k b_k}} \propto^+ \sigma_{x_k b_k} \sigma_{b_k x_{k+1}},$$

Therefore, the sign comparison holds and Theorem 6.5 follows. □

The next theorem extends Figure 6.8 to allow the conditionate Z to have any number of descendants on each collider x_k. In particular, we look at polytrees with the structure seen in Figure 6.9.

Theorem 6.6 *Consider the DAG in Figure 6.9. For $k = 1, ..., n$, let $Z_k = \{z_{k1}, ..., z_{k,n_k}\}$,*

$\mathcal{Z}_k = \cup_{i=1}^{k} Z_i$ *and* $\mathcal{Z}_k^* = \mathcal{Z}_k \cup \{z_{k+1,1}, z_{k+2,1}, ..., z_{n,1}\}$. *Then we have*

$$\sigma_{ac|\mathcal{Z}_n} \propto^+ \sigma_{ac|\mathcal{Z}_0^*}$$

where $\mathcal{Z}_0^* = \{z_{11}, ..., z_{n,1}\}$, $\mathcal{Z}_n^* = \mathcal{Z}_n$.

Proof: Let $\mathcal{Z}_{k+1}^{**} = \mathcal{Z}_{k+1}^* \backslash Z_{k+1}$. The proof is by induction. For $k = 1$, using Proposition 6.1 with $Z_1 \perp\!\!\!\perp ac|x_1\mathcal{Z}_1^{**}$, since $a \perp\!\!\!\perp c|\mathcal{Z}_1^{**}$, it is straightforward that

$$\begin{aligned}\sigma_{ac|Z_1^*} &= \sigma_{ac|\mathcal{Z}_1^{**}} - \Sigma_{aZ_1|\mathcal{Z}_1^{**}}\Sigma^{-1}_{Z_1Z_1|\mathcal{Z}_1^{**}}\Sigma_{Z_1c|\mathcal{Z}_1^{**}} \\ &= -\frac{\sigma_{ax_1|\mathcal{Z}_1^{**}}\sigma_{x_1c|\mathcal{Z}_1^{**}}}{\sigma^2_{x_1x_1|\mathcal{Z}_1^{**}}}\Sigma_{x_1Z_1|\mathcal{Z}_1^{**}}\Sigma^{-1}_{Z_1Z_1|\mathcal{Z}_1^{**}}\Sigma_{Z_1x_1|\mathcal{Z}_1^{**}}.\end{aligned}$$

Since $\Sigma^{-1}_{Z_1Z_1|\mathcal{Z}_1^{**}}$ is positive definite, using Proposition 6.1 with $z_1 \perp\!\!\!\perp ac|x_1\mathcal{Z}_1^{**}$, we get

$$\sigma_{ac|\mathcal{Z}_1^*} \propto^+ -\sigma_{ax_1|\mathcal{Z}_1^{**}}\sigma_{x_1c|\mathcal{Z}_1^{**}}. \tag{6.6.13}$$

$$\tag{6.6.14}$$

Using Proposition 6.1 with $z_1 \perp\!\!\!\perp ac|x_1\mathcal{Z}_1^{**}$, from (6.6.13) and the fact that $a \perp\!\!\!\perp c|\mathcal{Z}_1^{**}$, we get

$$\begin{aligned}\sigma_{ac|\mathcal{Z}_0^*} &= -\frac{\sigma_{az_1|\mathcal{Z}_1^{**}}\sigma_{z_1c|\mathcal{Z}_1^{**}}}{\sigma_{z_1z_1|\mathcal{Z}_1^{**}}} \\ &= -\frac{\sigma_{ax_1|\mathcal{Z}_1^{**}}\sigma^2_{z_1x_1|\mathcal{Z}_1^{**}}\sigma_{x_1c|\mathcal{Z}_1^{**}}}{\sigma_{x_1x_1|\mathcal{Z}_1^{**}}\sigma_{z_1z_1|\mathcal{Z}_1^{**}}} \\ &\propto^+ \sigma_{ac|\mathcal{Z}_1^*}.\end{aligned}$$

Suppose that it holds $\sigma_{ac|\mathcal{Z}_k^*} \propto^+ \sigma_{ac|\mathcal{Z}_0^*}$. We want to show that $\sigma_{ac|\mathcal{Z}_{k+1}^*} \propto^+ \sigma_{ac|\mathcal{Z}_0^*}$. Using Proposition 6.1 with $ac \perp\!\!\!\perp Z_{k+1}|Z_{k+1}^{**}x_{k+1}$, we have

$$\sigma_{ac|\mathcal{Z}_{k+1}^*} = \sigma_{ac|\mathcal{Z}_{k+1}^{**}} - \Sigma_{aZ_{k+1}|\mathcal{Z}_{k+1}^{**}}\Sigma^{-1}_{Z_{k+1}Z_{k+1}|\mathcal{Z}_{k+1}^{**}}\Sigma_{Z_{k+1}c|\mathcal{Z}_{k+1}^{**}}$$

$$= -\frac{\sigma_{ax_{k+1}|\mathcal{Z}^{**}_{k+1}}\sigma_{x_{k+1}c|\mathcal{Z}^{**}_{k+1}}}{\sigma^2_{x_{k+1}x_{k+1}|\mathcal{Z}^{**}_{k+1}}}\Sigma_{x_{k+1}Z_{k+1}|\mathcal{Z}^{**}_{k+1}}\Sigma^{-1}_{Z_{k+1}Z_{k+1}|\mathcal{Z}^{**}_{k+1}}\Sigma_{Z_{k+1}x_{k+1}|\mathcal{Z}^{**}_{k+1}}.$$

Since $\Sigma^{-1}_{Z_{k+1}Z_{k+1}|\mathcal{Z}^{**}_{k+1}}$ is positive definite, we get

$$\sigma_{ac|\mathcal{Z}^{*}_{k+1}} \propto^{+} -\sigma_{ax_{k+1}|\mathcal{Z}^{**}_{k+1}}\sigma_{x_{k+1}c|\mathcal{Z}^{**}_{k+1}} \tag{6.6.15}$$

Obviously, $\mathcal{Z}^{**}_{k+1} = Z^{*}_{k}\backslash z_{k+1}$, therefore using Proposition 6.1 with $ac \perp\!\!\!\perp z_{k+1}|\mathcal{Z}^{**}_{k+1}x_{k+1}$, we have

$$\begin{aligned}
\sigma_{ac|\mathcal{Z}^{*}_{k}} &= -\frac{\sigma_{az_{k+1}|\mathcal{Z}^{**}_{k+1}}\sigma_{z_{k+1}c|\mathcal{Z}^{**}_{k+1}}}{\sigma_{z_{k+1}z_{k+1}|\mathcal{Z}^{**}_{k+1}}} \\
&= -\frac{\sigma_{ax_{k+1}|\mathcal{Z}^{**}_{k+1}}\sigma_{x_{k+1}c|\mathcal{Z}^{**}_{k+1}}\sigma^2_{x_{k+1}z_{k+1}|\mathcal{Z}^{**}_{k+1}}}{\sigma^2_{x_{k+1}x_{k+1}|\mathcal{Z}^{**}_{k+1}}\sigma_{z_{k+1}z_{k+1}|\mathcal{Z}^{**}_{k+1}}} \\
&\propto^{+} -\sigma_{ax_{k+1}|\mathcal{Z}^{**}_{k+1}}\sigma_{x_{k+1}c|\mathcal{Z}^{**}_{k+1}}.
\end{aligned} \tag{6.6.16}$$

Therefore, using (6.6.15) and (6.6.16), we conclude that

$$\sigma_{ac|\mathcal{Z}^{*}_{k+1}} \propto^{+} \sigma_{ac|\mathcal{Z}^{*}_{k}} \propto^{+} \sigma_{ac|\mathcal{Z}^{*}_{0}}.$$

□

A similar proof can be extended to include conditionates $\mathcal{U}$ and $\mathcal{W}$ defined in Theorem 6.5.

Theorem 6.6 shows that the sign of $\sigma_{ac|Z}$ for any conditionate Z depends on the colliders on the path. In particular for two conditionates Z_1 and Z_2, $\sigma_{ac|Z_1} \propto^{+} \sigma_{ac|Z_2}$. This leads to the following corollary.

Corollary 6.1 *Consider a Gaussian polytree. Let $Z_1, Z_2 \subseteq V\backslash {}_a\pi_c$ such that $\forall z \in Z^{(1)} \cup Z^{(2)}$, ${}_z\pi_{\mathfrak{n}(z)}$ does not have a collider. Define*

$$\begin{aligned}
\mathcal{Z}^{(1)}_i &= \{z \in Z_i : \text{ at least one of the path } {}_a\pi_z, {}_c\pi_z \text{ has a collider } \} \\
\mathcal{Z}^{(2)}_i &= \{z \in Z_i : {}_a\pi_c, {}_a\pi_z, {}_c\pi_z \text{ do not have a collider at } \mathfrak{n}(z)\}
\end{aligned}$$

$$\mathcal{Z}_i^{(3)} = \{z \in Z_i : \ \textit{Only } {}_a\pi_c \textit{ has a collider at } \mathfrak{n}(z), \\ \textit{but } {}_a\pi_z, {}_c\pi_z \textit{ do not have a collider at } \mathfrak{n}(z).\}$$

If all of the conditions below are satisfied. That is,

(1) $Z_2^{(2)} \perp\!\!\!\perp {}_a\pi_c | Z_1^{(2)}$,

(2) $Z_1^{(1)} \perp\!\!\!\perp {}_a\pi_c | Z_2^{(1)}$,

(3) $Z_1^{(3)} \perp\!\!\!\perp {}_a\pi_c | Z_2^{(3)}$.

Then, exactly one of the two statements below holds.

(1) $\rho_{ac|Z_2} \geq \rho_{ac|Z_1} \geq 0$.

(2) $\rho_{ac|Z_2} \leq \rho_{ac|Z_1} \leq 0$.

Proof: From the results in Chaudhuri [2005] Proposition 2, page 23. We have $\rho^2_{ac|Z_1} \leq \rho^2_{ac|Z_2}$. From Theorem 6.5 and 6.6, it is clear that $\rho_{ac|Z_1}$ and $\rho_{ac|Z_2}$ have the same sign. Therefore, Corollary 6.1 follows.

□

The results are extensions to the results of key cases discussed in section 6.3. These results can be used in high-dimensional graphical model selection. In particular, these results specify bounds of deviation from faithfulness of the graph to its underlying distribution. We refer to Uhler et al. [2013] and Lin et al. [2012] for further details.

In the next section, we show that almost qualitative comparison of partial correlations lead to necessary and sufficient conditions for observations generated from a single factor model.

6.7 Application to Single Factor Model

Single factor models or star decomposable models are popular in psychometry, statistical finance, among others. In this model one assumes that the observations are influenced by one hidden variable. The observations are marginally dependent but conditionally independent given the hidden variable concerned. Since the hidden factor is

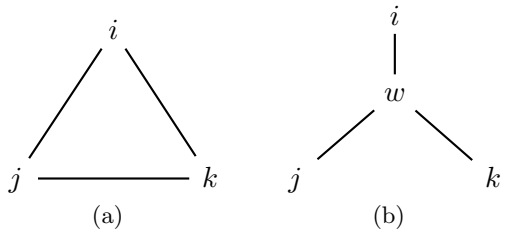


Figure 6.10 Figure 6.10(b) is the star model studied by Xu and Pearl [1989] while Figure 6.10(a) is the model observed using the marginal distribution

not observed the task is then to identify if the model is a single factor model from the observations.

An example of a single factor model is shown in Figure 6.10(b). If w is not observed, and i, j and k are only observed, the marginal distribution looks like Figure 6.10(a). Thus, neither the covariance matrix or the precision matrix shows any zero. Also, the standard penalization or methods which find zeros in covariance or precision matrix cannot be used to identify Figure 6.10(b) from Figure 6.10(a). Necessary and sufficient conditions to identify single factor model from the observed data has been huge interest to statisticians.

Such necessary and sufficient conditions have been studied by several authors before. Notable among them are Xu and Pearl [1989], Paul A. Bekker [1987]. Kuroki and Cai [2006]. They study the necessary and sufficient condition when the observations come from a single factor model but they are observed only for a strata of some variable.

A more general result is presented later. The next proposition is due to Anderson and Rubin [1956]. We present this here for completeness.

Proposition 6.2 *A four dimensional Gaussian distribution factors according to the graph in Figure 6.10(b) iff for all* $x, y, z \in \{i, j, k\}$, $x \neq y \neq z$, $\rho_{xw}^2 = \rho_{xy}\rho_{xz}/\rho_{yz}$.

Proof: ($\Rightarrow$) Clearly, in the graph G in Figure 6.10(b) for any $x \neq y \in \{i, j, k\}$, $x \perp\!\!\!\perp y \mid w$. Thus any Gaussian distribution factoring according to G would satisfy

$$\rho_{xy} = \rho_{xw}\rho_{yw}.$$

By substituting this expression we get

$$\rho_{ij}\rho_{ik}/\rho_{jk} = \rho_{iw}^2.$$

The proof for ρ_{jw}^2 and ρ_{kw}^2 are similar.
($\Leftarrow$) By assumption for all $x \neq y \neq z$, $0 \leq \rho_{xy}\rho_{xz}/\rho_{yz} \leq 1$. Now, for $x \neq y$, since

$$\rho_{xw}\rho_{yw} = \sqrt{\frac{\rho_{xy}^2\rho_{yz}\rho_{xz}}{\rho_{yz}\rho_{xz}}} = \rho_{xy},$$

it is straightforward that

$$\rho_{xy|w} = \rho_{xy} - \rho_{xw}\rho_{yw} = 0.$$

Thus, for any $x \neq y \in \{i, j, k\}$, $x \perp\!\!\!\perp y \mid w$, which implies that the distribution factors according to the graph in Figure 6.10(b). □

We now present a necessary and sufficient condition for three Gaussian random variable to be star decomposable based on the results presented in this section.

Theorem 6.7 *A necessary and sufficient condition for three random variables with a joint Gaussian distribution to be star-decomposable (see Figure 6.10(b) is that for all* $i, j, k \in \{1, 2, 3\}$, $i \neq j \neq k$*:*

(1) $\rho_{ij|k}^2 \leq \rho_{ij}^2$ *and*

(2) $\rho_{ij} \propto^+ \rho_{ij|k}$.

Proof: ($\Rightarrow$) If the joint Gaussian distribution is star decomposable, Theorem 6.1 and Chaudhuri [2013, Theorem 2] show that the first statement holds. Furthermore, $\rho_{ij|k}^2$ has the same sign as $\rho_{ij} - \rho_{ik}\rho_{jk}$. From the star decomposition $\rho_{ik}\rho_{jk} = \rho_{ij}\rho_{kw}^2$. So since $0 \leq \rho_{kw}^2 \leq 1$, we have

$$\rho_{ij} - \rho_{ik}\rho_{jk} = \rho_{ij}(1 - \rho_{kw}^2) \propto^+ \rho_{ij}.$$

($\Leftarrow$) From Xu and Pearl [1989, Theorem 2] and Proposition 6.2 above, it suffices to show that $0 \leq (\rho_{ik}\rho_{jk}/\rho_{ij}) \leq 1$.

First note that,

$$\rho^2_{ij|k} = \frac{(\rho_{ij} - \rho_{ik}\rho_{jk})^2}{\left\{(1-\rho^2_{ik})(1-\rho^2_{jk})\right\}} \leq \rho^2_{ij}.$$

This implies $(\rho_{ij} - \rho_{ik}\rho_{jk})^2 \leq \rho^2_{ij}$. Therefore,

$$0 \leq \rho^2_{ik}\rho^2_{jk} \leq 2\rho_{ij}\rho_{ik}\rho_{jk}.$$

Thus $\rho_{ij}\rho_{ik}\rho_{jk} \geq 0$ and $0 \leq \rho_{ik}\rho_{jk}/\rho_{ij} \leq 2$.

Now from the sign conditions we note that $\rho_{ij|k}$ has the same sign as $(\rho_{ij} - \rho_{ik}\rho_{jk})$. Now if $\rho_{ij} - \rho_{ik}\rho_{jk} \geq 0$, then $\rho_{ij} \geq 0$ and $\rho_{ik}\rho_{jk}/\rho_{ij} \leq 1$. On the other hand, if $\rho_{ij} - \rho_{ik}\rho_{jk} \leq 0$, then $\rho_{ij} \leq 0$ and still $\rho_{ik}\rho_{jk}/\rho_{ij} \leq 1$. Now by the same argument as Xu and Pearl [1989, Theorem 2] the conclusion follows. □

The necessary and sufficient condition for four or more observations follow from Theorem 6.7. We provide an alternative to Paul A. Bekker [1987].

Theorem 6.8 *Suppose X_1, X_2, ..., X_n, $n \geq 4$ are jointly Gaussian with a positive definite covariance matrix. Then a necessary and sufficient condition that they are star-decomposable is that for $i, j, k, l \in \{1, \ldots, n\}$, $i \neq j \neq k \neq l$:*

(1) $\rho^2_{X_iX_j|X_k} \leq \rho^2_{X_iX_j}$,

(2) $\rho_{X_iX_j} \propto^+ \rho_{X_iX_j|X_k}$ *and*

(3) $\rho_{X_iX_k}\rho_{X_jX_l} = \rho_{X_iX_l}\rho_{X_jX_k}$.

Proof: Follows directly from Theorem 6.7, Xu and Pearl [1989] and the assumption that the covariance is positive definite. □

Condition 3 in Theorem 6.8 is the tetrad condition. This condition excludes the graphs of the form shown in Figure 6.11. If the correlation matrix is positive definite, the tetrad condition will be satisfied. There has been a lot of work on matrices satisfying tetrad conditions. For details, refer to Spirtes et al. [2000].

Theorem 5.7 and 5.8 state the necessary and sufficient conditions in terms of properties of correlation matrix in the population. In practice, all these conditions have to be

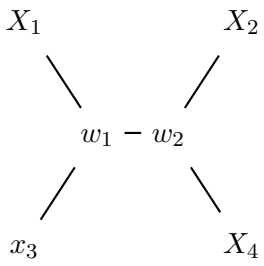


Figure 6.11 The graph above satisfy condition 1 and 2 of Theorem 6.8, but not condition 3

tested from the available data. The optimal testing procedures for such null hypotheses are not known. However, these results can readily be used on an exploratory basis.

6.8 Discussion

In this chapter we showed that the partial correlation and regression coefficients of a Gaussian random vector may not be compared qualitatively. However, under certain condition the comparison can be almost qualitative. In most cases, these conditions are determined by the covariance between the correlates, conditionates and a few other components. Thus the signs can be easily determined from the data without observing the whole vector and qualitative comparisons can be made.

We applied our results in characterizing single factor or star decomposable models. We also provide rules for comparison on the trees and a class of polytrees. Our rules can be applied to bigger classes of graphical Markov models. This may facilitate models selection of such models.

Bibliography

T. W. Anderson and H. Rubin. Statistical inference in factor analysis. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, 1954–1955, vol. V*, pages 111–150, Berkeley and Los Angeles, 1956. University of California Press.

T. M. Apostol. *Mathematical Statistics*. Narosa Publishing House, 1997.

L. Breiman, J. Friedman, R. Olshen, and C. Stone. *Classification and Regression Trees*. Wadsworth and Brooks, Monterey, CA, 1984.

C. Brito and J. Pearl. Generalised instrumental variables. *UAI 2002*, pages 85–93, 2002.

S. Chaudhuri. *Using the Structure Of d-connecting Paths As a Qualitative Measure of the Strength of Dependence*. PhD thesis, Seattle, WA, USA, 2005. AAI3183347.

S. Chaudhuri. Qualitative inequalities for squared partial correlations of a gaussian random vector. Technical Report 1/2013, Department of Statistics and Applied Probability, National University of Singapore, 2013.

S. Chaudhuri and T. S. Richardson. Using the structure of d-connecting paths as a qualitative measure of the strength of dependence. In *Uncertainty in Artificial Intelligence*, pages 116–123. Morgan Kaufmann Publishers, 2003.

S. Chaudhuri and G. L. Tan. On qualitative comparison of partial regression coefficients for gaussian graphical markov models. In Marlos A. G. Viana and Henry P Wynn, editors, *Algebraic methods in Statistics and Probability II*, volume 516 of *Contemporary Mathematics*, pages 125–133. American Mathematical Society, 2010.

A. P. Dempster. Covariance selection. *Biometrics*, 28(1):157–175, 1972.

M. Drton and M. D. Perlman. Model selection for gaussian concentration graphs. *Biometrika*, 91(3):591–602, 2004.

M. Drton and M. D. Perlman. A SINful approach to Gaussian graphical model selection. *J. Statist. Plann. Inference*, 138(4):1179–1200, 2008.

B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least angle regression. *Ann. Statist.*, 32(2):407–499, 2004.

J. Fan and R. Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.*, 96(456):1348–1360, 2001.

I. E. Frank and J. H. Friedman. A Statistical View of Some Chemometrics Regression Tools. *Technometrics*, 35(2):109–135, 1993.

J. H. Friedman, T. Hastie, H. Höfling, and R. Tibshirani. Pathwise coordinate optimization. *Ann. Appl. Stat.*, 1(2):302–332, 2007.

J. H. Friedman, T. Hastie, and R. Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 2008.

J. H. Friedman, T. Hastie, and R. Tibshirani. Applications of the lasso and grouped lasso to the estimation of sparse graphical models. 2010.

C. J. Geyer. On the asymptotics of convex stochastic optimization. *Unpublished manuscript.*, 1996.

S. Greenland. Quantifying biases in causal models: classical confounding versus collider-stratification bias. *Epidemiology*, 14:300–306, 2003.

T. Hastie, R. Tibshirani, and J. H. Friedman. *The elements of statistical learning*. Springer Series in Statistics. Springer, New York, second edition, 2009. ISBN 978-0-387-84857-0. Data mining, inference, and prediction.

S. Holm. A simple sequentially rejective multiple test procedure. *Scand. J. Statist.*, 6 (2):65–70, 1979. ISSN 0303-6898.

D. A. Holton and J. Sheehan. *The Petersen graph*, volume 7. Cambridge University Press, Cambridge, 1993.

M. Kendall and A. Stuart. *The Advanced Theory of Statistics, vol. 2, Inference and Relationship*. Macmillan Publishing Co., Inc., 1979.

K. Knight and W. Fu. Asymptotics for lasso-type estimators. *Annals of Statistics*, 28: 1356–1378, 2000.

M. Kuroki and Z. Cai. On recovering a population covariance matrix in the presence of selection bias. *Biometrika*, 93(3):601–611, 2006.

O. Laule, A. Fürholz, H. S. Chang, T. Zhu, X. Wang, P. B. Heifetz, W. Gruissem, and M. Lange. Crosstalk between cytosolic and plastidial pathways of isoprenoid biosynthesis in arabidopsis thaliana. *Proc Natl Acad Sci U S A*, 100(11):6866–71, 2003.

S. L. Lauritzen. *Graphical Models*. Oxford University Press, 1996.

S. Lin, C. Uhler, B. Sturmfels, and P. Bühlmann. Hypersurfaces and their singularities in partial correlation testing. *ArXiv e-prints*, September 2012.

H. Linhart and W. Zucchini. *Model selection*. Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics. John Wiley & Sons Inc., New York, 1986. ISBN 0-471-83722-9.

K. V. Mardia, J. T. Kent, and J. M. Bibby. *Multivariate Analysis*. Academic Press, 1979.

N. Meinshausen and P. Bülmann. High dimensional graphs and variable selection with the lasso. *Annals of Statistics*, 34:1436–1462, 2006.

MIM. Mim 3.1 student version, jun 2009. URL `[http://www.hypergraph.dk]`.

M. R. Osborne, B. Presnell, and B. A. Turlach. A new approach to variable selection in least squares problems. *IMA J. Numer. Anal.*, 20(3):389–403, 2000.

J. de Leeuw Paul A. Bekker. The rank of reduced dispersion matrices. *Psychometrika*, 52:125–135, 1987.

S. E. Payne. Finite generalized quadrangles: a survey. In *Proceedings of the International Conference on Projective Planes (Washington State Univ., Pullman, Wash., 1973)*, pages 219–261. Washington State Univ. Press, Pullman, Wash., 1973.

J. Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2000.

J. Peng, P. Wang, N. Zhou, and J. Zhu. Partial correlation estimation by joint sparse regression models. *Journal of the American Statistical Association*, 104(486):735–746, 2009.

M. Pourahmadi. Maximum likelihood estimation of generalised linear models for multivariate normal covariance matrix. *Biometrika*, 87:425–435, 2000.

M. Rodriguez-Concepcion and A. Boronat. Elucidation of the Methylerythritol Phosphate Pathway for Isoprenoid Biosynthesis in Bacteria and Plastids. A Metabolic Milestone Achieved through Genomics. *Plant Physiol.*, 130(3):1079–1089, 2002.

M. Rodriguez-Concepcion, O. Fores, J.F. Martinez-Garcia, V. Gonzalez, M.A. Phillips, A. Ferrer, and A. Boronat. Distinct light-mediated pathways regulate the biosynthesis and exchange of isoprenoid precursors during arabidopsis seedling development. *Plant Cell*, 16(1):144–56, 2004.

A. Shojaie and G. Michailidis. Penalized likelihood methods for estimation of sparse high-dimensional directed acyclic graphs. *Biometrika*, 97(3):519–538, 2010.

Z. Šidák. Rectangular confidence regions for the means of multivariate normal distributions. *J. Amer. Statist. Assoc.*, 62:626–633, 1967.

T. P. Speed and H. T. Kiiveri. Gaussian Markov distributions over finite graphs. *Ann. Statist.*, 14(1):138–150, 1986.

P. Spirtes, C. Glymour, and R. Scheines. *Causation, Prediction and Search*. MIT Press, 2000.

R. Tibshirani. Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B*, 58(1):267–288, 1996.

C. Uhler, G. Raskutti, P. Bühlmann, and B. Yu. Geometry of the faithfulness assumption in causal inference. *Ann. Statist.*, 41(2):436–463, 2013.

L. Vandenberghe, S. Boyd, and S. P. Wu. Determinant maximization with linear matrix inequality constraints. *SIAM J. Matrix Anal. Appl.*, 19(2):499–533, 1998.

T. Verma and J. Pearl. Equivalence and synthesis of causal models. In *Uncertainty in Artificial intelligence*, pages 220–227, 1990.

J. Whittaker. *Graphical Models in Applied Multivariate Statistics*. Wiley, 1990.

A. Wille, P. Zimmermann, E. Vranova, A. Fürholz, O. Laule, S. Bleuler, L. Hennig, A. Prelić, P. von Rohr, L. Thiele, E. Zitzler, W. Gruissem, and P. Bühlmann. Sparse Graphical Gaussian Modeling of the Isoprenoid Gene Network in Arabidopsis thaliana. *Genome Biol*, 5(11):R92, 2004.

L. Xu and J. Pearl. Structuring causal tree models with continuous variables. In *Uncertainty in Artificial Intelligence*, pages 170–178. Morgan Kaufmann Publishers, 1989.

M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 68(1):49–67, 2006.

M. Yuan and Y. Lin. Model selection and estimation in the gaussian graphical model. *Biometrika*, 94(1):19–35, 2007.

P. Zhao and B. Yu. On model selection consistency of Lasso. *J. Mach. Learn. Res.*, 7: 2541–2563, 2006.

S. Zhou, P. Rütimann, M. Xu, and P. Bühlmann. High-dimensional covariance estimation based on Gaussian graphical models. *J. Mach. Learn. Res.*, 12:2975–3026, 2011. ISSN 1532-4435.

H. Zou. The adaptive lasso and its oracle properties. *J. Amer. Statist. Assoc.*, 101(476): 1418–1429, 2006. ISSN 0162-1459.

H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 67(5):768, 2005.